Vers plus de contrôle pour la synthèse de parole expressive

Damien Lolive

UNIVERSITÉ DE RENNES 1

École doctorale MATISSE

Habilitation à diriger des recherches
présentée par

Damien Lolive

Vers plus de contrôle pour la synthèse de parole expressive

soutenue publiquement le 29/11/2017

devant le jury composé de

Véronique Delvaux	Chercheuse qualifiée FNRS, Université de Mons	Rapportrice
Frédéric Béchet	Professeur, Aix Marseille Université	Rapporteur
Yves Laprie	Directeur de recherche CNRS, LORIA	Rapporteur
Philippe Martin	Professeur, Université Paris Diderot	Examinateur
François Goasdoué	Professeur, Université de Rennes 1	Examinateur

Table des matières

Introduction

Le présent document résume la majeure partie des travaux que j'ai menés depuis l'obtention de mon doctorat en 2008. Ces travaux ont été conduits à l'université de Rennes 1 et à l'IRISA, tout d'abord au sein de l'équipe Cordial puis de l'équipe Expression, dont je suis l'un des fondateurs. Le cadre de mes travaux est celui du traitement automatique de la parole avec pour application la synthèse de parole expressive. Dans ce cadre, les thématiques que j'aborde sont :

- la constitution de corpus en vue de la création de voix artificielles : cette dernière repose sur des corpus textuels annotés dont la création conditionne la qualité de la voix créée, cette thématique est donc une fondation nécessaire pour aborder les suivantes ;
- la caractérisation de l'expressivité sur un plan prosodique : il s'agit de comprendre ce qui fait l'expressivité d'une voix, de manière à pouvoir la reproduire en introduisant de nouvelles informations et de nouveaux processus dans la chaîne de synthèse de parole ;
- le traitement du langage naturel : le langage est présent à tous les niveaux en traitement de la parole de par les annotations nécessaires (texte, mots, phonèmes, style, prosodie, etc) ;
- la synthèse de la parole pour laquelle deux axes sont traités à travers les deux principaux types de systèmes (statistiques et par concaténation) ;
- l'évaluation des systèmes de synthèse de la parole.

Ces travaux se sont fortement élargis depuis ces dernières années avec l'aboutissement de la création d'une librairie de représentation des corpus de manière structurée, d'un prototype de moteur de synthèse par sélection d'unités, de la coordination du projet ANR Phorevox, de nouvelles collaborations et également l'obtention du financement d'un projet ANR JCJC, dont je suis le coordinateur.

Contexte général Pour les systèmes de synthèse de la parole à partir du texte, l'intelligibilité a été la préoccupation majeure au détriment du naturel de la voix. De nos jours, c'est ce dernier aspect qui devient important (Murray et Arnott 1996). Non seulement la qualité du signal sonore et la présence d'artéfacts influencent le naturel perçu de

la voix produite, mais beaucoup d'autres aspects liés à l'état d'esprit du locuteur, son état émotif, son intention lorsqu'il énonce le message, le contexte de son intervention, sont autant de paramètres qui confèrent à la voix tout son naturel.

Les limitations en termes de flexibilité et de qualité des moteurs de synthèse de la parole sont un frein au développement d'applications cruciales pour la société numérique de demain. Naturellement, le développement de ces applications représente un enjeu industriel important. Ainsi, les domaines de l'éducation, du divertissement, de l'assistance aux personnes sont des exemples pour lesquels un impact majeur est attendu. Voici en particulier quelques applications qui bénéficieraient d'avancées dans le domaine de la synthèse de parole expressive :

- Jeux vidéo : pour la diversification des voix dans les jeux vidéo, la création de voix dont la prosodie peut varier fortement suivant le contexte (personnage principal en difficulté ou en situation de victoire) ;
- Lecture de livres audio : fournir des voix expressives permettant de transmettre la sémantique correcte du texte par l'expressivité de la voix, prosodie pouvant varier fortement au cours de la lecture, imitation de personnages au cours de la lecture ;
- Assistance aux utilisateurs dans des dispositifs dédiés qui nécessitent une prosodie adaptée (par exemple : GPS, système d'alerte et de prévention, d'information) ;
- Assistance aux personnes handicapées ayant perdu l'usage de leur voix et utilisant un système de synthèse de la parole dans la vie quotidienne ;
- Logiciels éducatifs : dictée avec une prosodie en phase avec la tâche, apprentissage des langues, apprentissage du maniement d'un style d'élocution particulier.

De plus, par extension d'applications liées à l'apprentissage, une autre application potentielle est la préservation des langues. En effet, les systèmes de synthèse de la parole peuvent être utilisés comme outil de transmission du savoir et également comme outil de préservation du patrimoine. Certaines langues sous-dotées sont en voie de disparition. C'est par exemple le cas des langues régionales en France mais c'est aussi le cas d'un grand nombre de langues et de dialectes au niveau mondial (Australie, Amérique du nord, Afrique, etc.). La synthèse de parole pourrait être un outil afin de revitaliser certaines langues en danger. Dans ce cadre, la préservation de l'expressivité de la langue tout en s'adaptant à ses particularités par une prosodie adéquate est importante.

Compte-tenu des applications possibles, les retombées pour la société sont importantes. En particulier, l'accessibilité et le "confort" de personnes souffrant de handicap peut être grandement améliorée par un gain de qualité et de personnalisation des voix de synthèse. De même, une amélioration des processus d'apprentissage est possible en permettant à l'apprenant d'être plus autonome et de progresser à son rythme dans l'apprentissage d'une langue.

D'un point de vue technologique, à l'heure actuelle, trois catégories de systèmes de synthèse sont très utilisées : les systèmes de type statistiques (HTS), par réseaux de neurones profonds (DNN) et les systèmes par corpus. Les deux premiers types permettent de

s'adapter plus facilement à une expressivité particulière mais, en contrepartie, produisent une parole de moins bonne qualité que les systèmes de synthèse par corpus. Ces derniers, au contraire, offrent une très bonne qualité de parole synthétique mais sont beaucoup moins souples d'utilisation et, à l'heure actuelle, n'intègrent pas dans leur processus des moyens de contrôle pertinents de l'expressivité.

Expressivité et émotion Le terme expressivité est un terme pouvant regrouper divers éléments qui permettent de différencier une parole "neutre" d'une parole qui ne l'est pas (même si une parole neutre peut être considérée comme une forme d'expressivité). Ainsi, l'émotion, les styles d'élocution (dépendant du locuteur et de la tâche), l'intention du locuteur au moment de la prise de parole, peuvent être regroupés sous ce terme.

Pour ce qui nous intéresse, d'un point de vue fonctionnel, ces trois domaines font appel à des consignes placées en entrée du système (émotion, intention, style d'élocution), lesquelles vont influencer la génération de la sortie afin de refléter un certain style de parole. L'objectif premier dans ce contexte est d'identifier des descripteurs pertinents qui permettent de caractériser ces différents phénomènes, et ensuite de créer des modèles génératifs afin de créer une parole la plus naturelle possible qui prend en compte les phénomènes décrits.

Parmi ces trois phénomènes, le plus étudié et le plus difficile à définir est certainement l'émotion. Comme le présente Scherer 2005, la définition du concept d'émotion est un problème épineux dont plus d'une centaine de définitions ont pu être recensées. Il est tout de même intéressant de noter que comme le mentionne Vaudable 2012, la plupart des systèmes de reconnaissance des émotions considère un concept assez large incluant notamment émotion, attitude, humeur, etc.

Expressivité en synthèse de parole La prise en compte de l'émotion dans la génération d'un signal de parole a commencé à être étudiée au début des années 90, notamment par l'utilisation de modifications manuelles du signal de parole (Murray et Arnott 1996 ; Cahn 1990 ; Carlson 1992). Cahn 1990 propose ainsi une application de manipulation du signal de parole qui repose sur un système de synthèse de la parole à base de formants.

Par la suite, de nombreuses études ont permis des avancées sur la compréhension de la manifestation des émotions dans le signal de parole mais sans permettre encore la création de systèmes complets et générant un signal de parole de très haute qualité pour un large panel d'émotions et/ou d'expressivités (Z. Handley 2009). Selon (Murray et Arnott 1996), l'intelligibilité a été la préoccupation majeure au détriment du naturel de la voix. Dans (Rebordao et al. 2009), le même constat est présent, il n'existe pas à l'heure actuelle de système de synthèse complet qui prenne du texte en entrée, évalue son contenu expressif (affectif, émotionnel, intentionnel), et génère un signal de parole approprié. Selon (M. Schröder 2001), pour exprimer un nombre important d'émotions, soit les systèmes à base de règles doivent produire une parole plus naturelle, soit les

systèmes par concaténation doivent devenir plus flexibles. L'avènement de l'apprentissage profond permet une évolution du domaine en permettant l'apprentissage de modèles représentant directement le signal de parole. Ces dernières avancées laissent entrevoir une métamorphose du domaine de la synthèse de parole dans les prochaines années.

Dans tous les cas, la production d'un signal de parole à partir du texte fait intervenir des modèles prosodiques dont le rôle est de prédire à partir d'informations linguistiques, l'évolution des différents paramètres prosodiques telles que la fréquence fondamentale, l'intensité, les durées phonémiques et les pauses. Traditionnellement, ces modèles permettent de prendre en compte un seul type d'élocution qui jusqu'à récemment correspondait à un style neutre (lecture d'un texte). Les systèmes actuels manquent de contrôle sur l'expressivité que ce soit au niveau des traitements linguistiques, prosodiques mais également lors de l'étape de sélection des unités dans le cas de la synthèse par concaténation.

Pourtant, les enjeux de cette prise en compte de l'expressivité sont bien réels. On peut, par exemple, citer le domaine de l'apprentissage des langues assisté par ordinateur (CALL - Computer Assisted Language Learning) décrit dans (Z. Handley 2009). De manière plus générale, M. Eskenazi 2009 s'interroge également sur l'utilisation des technologies issues du traitement de la parole à l'heure actuelle et fait apparaître des besoins au niveau du support à l'apprentissage mais également dans le domaine des jeux vidéo.

Organisation du document L'objectif de mes travaux se concentre sur l'amélioration de la flexibilité du processus de synthèse de parole afin d'obtenir une meilleure gestion de l'expressivité. Ainsi, le présent document présente mes contributions au domaine de la synthèse de la parole à travers quatre axes complémentaires. Le chapitre 1, *Outils et méthodologies d'évaluation*, aborde le développement d'outils pour la représentation de corpus de parole ainsi que mes travaux autour de l'évaluation des systèmes de synthèse. Pour le premier point, une description de la librairie Roots est effectuée afin de montrer comment cette librairie permet la représentation structurée et cohérente des informations nécessaires à des travaux dans le domaine du traitement de la parole. La majorité des travaux présentés ensuite s'appuient sur cette librairie. Le second point concerne l'évaluation des systèmes, qui est indéniablement un enjeu dans le domaine. Une nouvelle méthodologie qui repose sur une sélection des échantillons à évaluer y est présentée. Cette dernière est utilisée à plusieurs reprises dans les chapitres suivants.

Le chapitre 2, *Modélisation de la prononciation*, décrit ensuite une approche pour la modélisation de la prononciation et de ses variantes. La modélisation de la prononciation est vue sous l'angle de l'adaptation par rapport à une prononciation de référence de manière à refléter une prononciation spécifique, par exemple à un style ou les habitudes d'un locuteur. Cette méthodologie est appliquée à la parole spontanée ainsi qu'au contexte spécifique d'une voix de synthèse. Les résultats montrent que la méthode est efficace à la fois pour adapter la prononciation à un style particulier et également réduire

les écarts entre prononciation canonique et prononciation spécifique à un locuteur.

Le chapitre 3, *Styles de parole pour la synthèse*, regroupe les travaux liés à l'étude des styles de parole en vue d'une adaptation des modules de prédiction de la prosodie pour la synthèse de styles spécifiques. Dans une première partie, quatre styles sont étudiés et comparés afin d'établir des règles les caractérisant applicables pour la synthèse. La deuxième partie de ce chapitre est quant-à-elle consacrée à l'étude du rythme avec une comparaison entre parole naturelle et parole synthétique. Enfin, une application à la génération de dictées est présentée et évaluée, montrant l'applicabilité de la synthèse des contextes précis, comme l'apprentissage des langues.

Enfin, le chapitre 4, *Adaptation du moteur de synthèse*, présente le moteur de synthèse de l'équipe Expression et ses évolutions. Notamment, l'intégration de contraintes lors du processus de sélection d'unités y est discutée à travers une étude sur l'intégration d'une consigne sur les durées phonémiques. Des évaluations du système sont également présentées ainsi que les deux participations de l'équipe au challenge de synthèse de parole Blizzard. Ces participations montrent la maturité du système développé et offre une perspective intéressante du travail mené sur le plan international.

Chapitre 1

Outils et méthodologies d'évaluation

*Les travaux présentés dans ce chapitre ont été conduits principalement à l'IRISA dans l'équipe de recherche Expression. Pour ces travaux, j'ai contribué sur le plan scientifique de manière importante mais également sur le plan organisationnel en coordonnant une partie des travaux dans le cadre du projet ANR Phorevox. Les publications suivantes en sont le résultat : (*Barbot, Barreaud *et al. 2011 ;* Boëffard, Laure *et al. 2012 ;* Boëffard, Charonnat *et al. 2012 ;* Chevelu, Lecorvé *et* Lolive *2014a ;* Chevelu, Lecorvé *et* Lolive *2014b ;* Lolive, Barbot *et* Boëffard *2009 ;* Chevelu *et al. 2015 ;* Chevelu *et* Lolive *2015 ;* Chevelu *et al. 2016).*

La construction des corpus et l'évaluation subjective des systèmes sont deux tâches cruciales pour faire avancer la recherche dans le domaine du traitement automatique de la parole. En effet, une gestion efficace et cohérente des données permet d'utiliser des corpus avec des volumes importants, de diversifier les contextes d'évaluation, de favoriser la reproductibilité des expérimentations. De plus, les méthodologies d'évaluation sont un enjeu important de la recherche en synthèse de parole. Le destinataire de parole synthétique, un être humain, est le seul qui soit apte à juger de la qualité de tel ou tel système. Cependant les méthodologies utilisées la plupart du temps, souvent en raison de la difficulté et du coût des évaluations, possèdent des biais.

Dans ce chapitre, les travaux que j'ai menés autour de ces deux problématiques sont présentés. Tout d'abord, ma contribution pour la représentation des corpus est introduite. Dans un deuxième temps, une méthodologie d'évaluation se concentrant sur les différences entre systèmes est présentée.

1.1 Représentation des corpus

La construction de corpus de parole est une étape cruciale pour tout système de synthèse de la parole à partir du texte. L'usage de modèles statistiques nécessite aujourd'hui l'utilisation de corpus de très grande taille qui doivent être enregistrés, transcrits, annotés et segmentés afin d'être exploitables. La variété des corpus nécessaires aux applications actuelles (contenu, style, etc.) rend l'utilisation de ressources audio disponibles, comme les livres audio, très attrayante. Ces corpus de parole sont annotés d'informations situées sur des plans différents allant le plus souvent de l'acoustique à la linguistique. Il est indispensable de disposer de structures de données adaptables permettant de représenter tous ces niveaux de description de manière cohérente et structurée.

1.1.1 Contexte

De nombreux outils de traitement automatique de la parole et du langage naturel permettent aujourd'hui d'annoter des documents écrits et oraux. Cependant, cette richesse logicielle conduit à une grande diversité de formats de fichiers et de types d'information. Du fait de cette hétérogénéité, le développement de processus de traitements complexes nécessite souvent de convertir et d'aligner, parfois de manière répétée, de nombreuses informations.

Nous décrivons ici une solution en réponse à ce problème de description cohérente faisant intervenir des séquences d'éléments, accompagnées de relations algébriques entre ces séquences. Ces dernières permettent notamment un passage rapide et aisé d'un niveau de représentation à un autre. Au delà des relations natives proposées entre séquences, nous proposons un mécanisme de composition algébrique permettant d'expliciter les mises en relation entre des éléments portés par des séquences arbitraires.

Pour pallier ce problème, nous présentons ensuite Roots, un outil libre dédié à la manipulation homogène de données séquentielles annotées. Roots est écrit en C++ et dispose d'une interface de programmation (API) dans plusieurs langages. Il est rapide et facile à prendre en main.

Toutefois, d'autres outils permettent le traitement de données annotées. Parmi eux, on peut citer le système GATE (Cunningham et al. 2002) qui propose d'interconnecter des flux d'annotations pour le développement de briques de TAL ; la boîte à outils NXT proposée pour la gestion de corpus multimodaux (Carletta et al. 2005 ; Calhoun et al. 2010) ou UIMA (Ferrucci et Lally 2004 ; Ferrucci, Lally et al. 2006) qui propose des normes de génie logiciel pour la gestion de données non structurées. Cependant, contrairement à ces approches, la philosophie de Roots est de laisser à l'utilisateur le soin de parcourir les données comme bon lui semble, de permettre la construction de ponts entre formats à moindre coût, et d'effectuer du prototypage rapide nécessaire à un travail de recherche.

ROOTS se rapproche des travaux réalisés dans le cadre du système de synthèse de la parole Festival (Alan W. BLACK et al. 2002). Ce système s'appuie sur le formalisme HRG, pour *Heterogenous Relation Graphs*, qui vise à offrir un mode de représentation unique des différents niveaux d'informations intervenant dans un système de synthèse (TAYLOR, Alan W. BLACK et CALEY 2001). Toutefois, ROOTS se distingue d'HRG car ROOTS est indépendant de tout outil et dispose par ailleurs d'une véritable interface de programmation (API), en C++ et en Perl.

1.1.2 Librairie ROOTS

Comme le résume la figure 1.1, les données sont structurées hiérarchiquement dans ROOTS. Fondamentalement, les données sont modélisées comme des séquences d'items. Ces items sont typés, par exemple, en mots, en graphèmes, en classes d'entités nommées, etc., et peuvent donc représenter différents niveaux d'annotations des mêmes données comme autant de séquences de types différents. Les correspondances entre items des différentes séquences sont définies comme des relations algébriques représentées sous forme de matrices. L'ensemble de ces relations permet de produire un graphe non orienté dont les nœuds sont des items et les arêtes sont dérivées des relations. La composition des relations algébriques qui forment le plus court chemin entre deux séquences permet de trouver des correspondances entre items même en l'absence de relation directe. L'exemple d'une séquence de mots annotée en POS et en phonèmes est donnée en figure 1.2. Plus de détail sur le calcul des relations est donné dans (BARBOT, BARREAUD et al. 2011).

Les séquences d'un même contenu sont regroupées en énoncés, ou *utterances* dans la terminologie de ROOTS. Ils peuvent correspondre à n'importe quelle unité pertinente pour un domaine donné (un mot isolé, une phrase, un groupe de souffle, etc.). Rassemblés en une liste, les énoncés forment un corpus. Beaucoup de structures peuvent être modélisées ainsi : des paragraphes, des chapitres, des livres, des reportages télévisés, des brèves d'actualités, etc. Pour hiérarchiser ces structures, un corpus peut être divisé en sous-corpus, par exemple pour représenter un chapitre comme une liste de paragraphes. Ces divisions « verticales » (*cf.* figure 1.1) d'un corpus sont appelées *chunks* dans ROOTS.

Les corpus peuvent aussi être découpés « horizontalement » en couches d'informations, ou *layers*. Ces couches permettent, au besoin, d'isoler certaines annotations relevant d'un même niveau d'abstraction. Par exemple, dans le cadre d'un corpus de journaux télévisés, il peut être intéressant de découper le corpus en 3 couches : l'une pour les informations acoustiques et phonétiques (fréquences fondamentales, spectres, allophones, etc.) ; une autre regroupant les informations d'ordre linguistique (transcription orthographique, étiquetages en POS, arbres syntaxiques, etc.) ; une dernière pour des métadonnées (heure de diffusion, lieu d'un reportage, etc.).

ROOTS s'appuie sur une bibliothèque écrite en C++ et rassemble un ensemble d'utilitaires. La bibliothèque représente environ 33 000 lignes de code. Elle s'accompagne

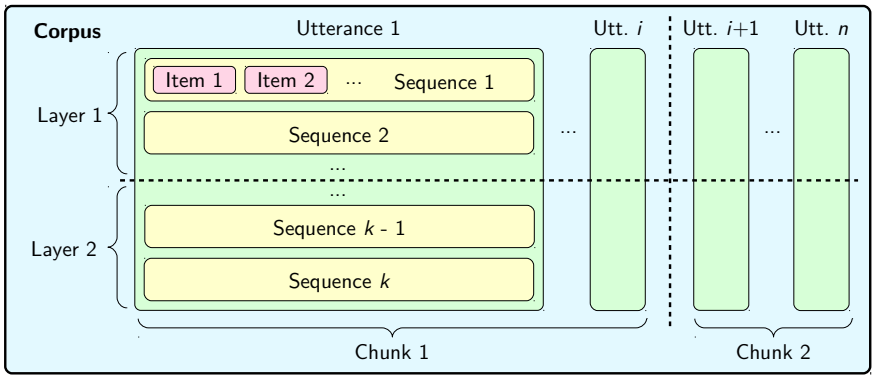

FIGURE 1.1 – *Organisation logique des données dans* ROOTS *(*CHEVELU, LECORVÉ *et* LOLIVE *2014a).*

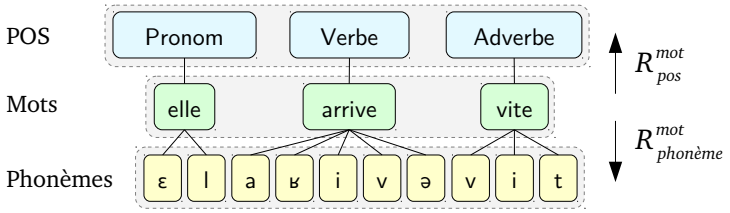

FIGURE 1.2 – *Exemple de 3 séquences liées par 2 relations,* R^{mot}_{pos} *et* $R^{mot}_{phonème}$ *(*CHEVELU, LECORVÉ *et* LOLIVE *2014a).*

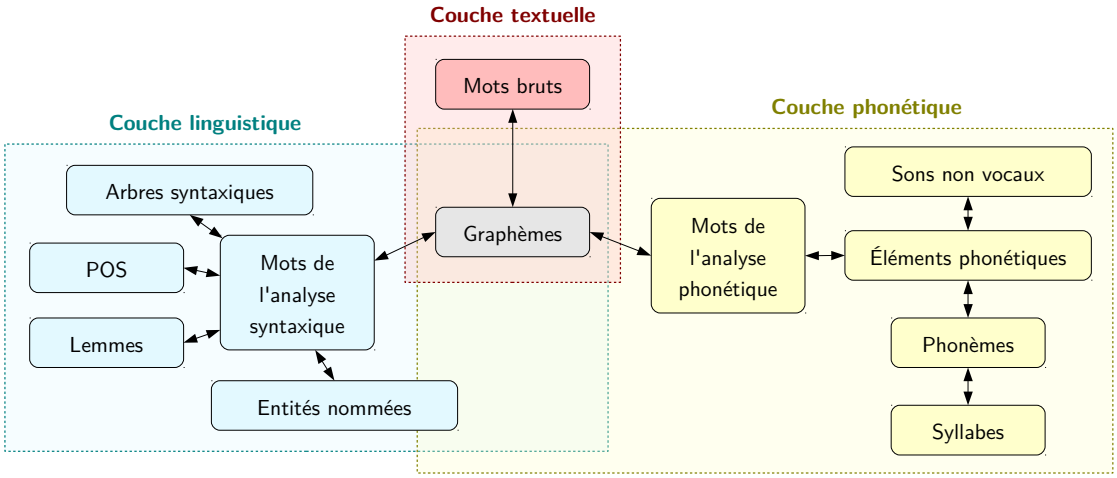

FIGURE 1.3 – *Vue schématique des relations entre séquences (*CHEVELU, LECORVÉ *et* LOLIVE *2014a).*

d'une API riche et largement documentée qui propose en particulier les fonctionnalités suivantes pour les différents éléments manipulés :

- **Items :** obtenir et définir les propriétés d'un item, comme le genre et le nombre

pour un POS ; obtenir les items liés dans d'autres séquences.

- **Relation :** tester, obtenir les items liés ; lier ou délier des items ;
- **Séquence :** ajout, suppression, extraction et mise à jour d'un ou plusieurs items.
- ***Utterance* :** ajout, suppression, extraction et mise à jour des séquences et relations internes ; obtenir des relations directes ou composées entre séquences ; combiner, scinder et sauvegarder.
- **Corpus :** ajout, suppression, extraction et mise à jour d'un énoncé ; gestion de la structure logique et du stockage ; sauvegarde et chargement.

Pour permettre le prototypage simple et rapide de programmes, cette API est aussi disponible pour les langages Perl et Python grâce au générateur automatique d'interface SWIG. Ainsi, ROOTS peut être porté à moindre frais vers de nombreux autres langages de programmation.

Des scripts utilitaires fondés sur l'API Perl sont disponibles pour les opérations les plus utiles (fusions et découpages de corpus, la recherche dans un corpus, la visualisation textuelle comme graphique des données...). D'autres scripts permettent également les imports et exports depuis et vers les outils de traitement de la parole HTK (YOUNG et al. 2009), Praat (BOERSMA 2002), WaveSurfer (SJÖLANDER et BESKOW 2000) et Transcriber (BARRAS et al. 2001). ROOTS permet d'ajouter simplement d'autres formats, par exemple ceux de type CSV.

Toutes les sources, la documentation et des tutoriaux sont disponibles [1] en ligne sur `https://bitbucket.org/lolived/roots/`.

1.1.3 Applications

Dans cette partie, deux exemples d'application sont abordés. Il est à noter que la maturité de la librairie est suffisante pour une utilisation quotidienne dans de nombreuses applications de recherche. En effet, tous les corpus que nous utilisons sont d'abord convertis au format ROOTS, ce qui permet d'unifier la gestion des données d'une application à l'autre (modélisation de la prononciation, synthèse de parole, étude de la prosodie, etc).

Construction d'un corpus de parole annoté à partir de livres-audio

Dans (BOËFFARD, LAURE et al. 2012 ; BOËFFARD, CHARONNAT et al. 2012), nous avons proposé une méthodologie de préparation et d'annotation de corpus de parole pour la synthèse de parole. Cette méthodologie s'appuie sur un ensemble d'outils existants, des livres-audio, ainsi que de la librairie ROOTS et a été appliquée avec succès sur 11 heures de parole extraites d'un livre audio. Une vérification manuelle sur une partie du corpus annoté a montré l'efficacité du procédé.

1. Sous licence *GNU Lesser General Public Licence* (LGPL) v3.0.

Type	Nombre (millions)
Mots (bruts)	94
Graphèmes	534
Mots (phonologiques)	112
Syllabes	134
Sons non vocaux	21
Phonèmes	309

Type	Nombre (millions)
Mots (linguistiques)	111
POS	111
Lemmes	111
Entités nommées	4
Arbres syntaxiques	6

TABLE 1.1 – *Nombre d'items automatiquement générés dans le corpus de livres numériques.*

Au cours de la première étape, l'alignement du texte et du son est réalisé : (1) découpage de l'enregistrement sur des pauses, (2) reconnaissance du texte associé à chaque fragment sonore par un système de reconnaissance automatique de la parole (ASR), (3) alignement entre le texte reconnu et le texte original.

Au cours de la deuxième étape, les descriptions textuelles et sonores sont fournies par l'objet ROOTS aux différents systèmes d'annotation qui donneront en retour leur propre analyse. Actuellement, les niveaux d'annotations utilisés sont les suivants : une extraction d'entités nommées, une analyse syntaxique, une analyse en POS (Part-Of-Speech), une segmentation phonétique et une extraction des proéminences prosodiques. Les analyses syntaxiques et grammaticales sont réalisées par des logiciels fournis par Synapse Développement, les analyses acoustiques sont réalisées à l'aide de nos propres outils. Les informations obtenues à chaque analyse sont intégrées au fichier ROOTS affinant ainsi la description du corpus et permettant d'établir de nouvelles relations entre les éléments des différentes annotations.

L'intérêt d'utiliser ROOTS est qu'il permet de conserver toutes les informations extraites alignées et cohérentes. L'utilisation du corpus s'en trouve ensuite facilitée pour de nombreuses applications.

Construction d'un corpus annoté de livres numériques

Nous avons conçu un processus non supervisé qui, à partir d'un texte brut en français, produit et relie les informations suivantes : graphèmes, POS, lemmes, entités nommées, arbres syntaxiques, phonèmes, sons non vocaux et syllabes. En utilisant les concepts de ROOTS et comme illustré par la figure 1.3, ce processus consiste à compléter une couche d'informations textuelles rudimentaires, composées des seules séquences de mots bruts et des graphèmes correspondants, avec une couche linguistique et une couche phonologique.

Ce processus d'annotation a été appliqué sur 1 300 livres numériques, libres et en français, collectés depuis http://www.ebooksgratuits.com. Le texte brut de ces livres a ainsi été

transformé en un corpus ROOTS dont les statistiques sont présentées dans la table 1.1. Ce corpus est réparti en 35 388 fichiers JSON qui occupent un total de 220 Go (environ 27 fichiers et 173 Mo par livre). Cette taille pourrait être largement réduite par compression puisque JSON est un format très verbeux. Le temps d'exécution du processus d'annotation a été de quinze heures et comprend principalement des appels aux outils externes comme le service en ligne de phonétisation.

1.2 Évaluation des systèmes de synthèse de parole

L'évaluation subjective des systèmes de synthèse de la parole est cruciale puisque l'objectif principal de ces systèmes est de produire un message destiné à des auditeurs humains. Cependant, pour être utiles, les évaluations subjectives nécessitent un ensemble de testeurs et d'échantillons significatifs, choisis en fonction du contexte d'utilisation de la synthèse.

Plusieurs types d'évaluations perceptives sont généralement utilisés. Parmi toutes ces méthodes, on peut distinguer des tests de préférence comme AB et ABX, des tests de pointage comme MOS (*Mean Opinion Score*), DMOS (*Degradation MOS*) ou plus récemment MUSHRA (*MUltiple Stimuli with Hidden Reference and Anchor*). Toutes ces méthodes ont le même objectif, à savoir le classement des systèmes selon certains critères subjectifs.

Dans cette partie, nous présentons une méthode d'évaluation des systèmes de synthèse dont l'objectif est d'améliorer la significativité des résultats.

1.2.1 Contexte

Dans la littérature, la plupart des propositions scientifiques sont évaluées à l'aide de tests perceptifs mais le nombre d'échantillons étudiés reste très limité. Par exemple, le défi Blizzard est composé de campagnes d'évaluation à grande échelle (PRAHALLAD, VADAPALLI et al. 2015 ; Simon KING et KARAISKOS 2016) mais ne comporte que quelques centaines de signaux. Ceci se retrouve dans d'autres travaux, parmi lesquels nous pouvons citer (SAINZ et al. 2014) avec 350 phrases, (GARCIA et al. 2006) avec 7 phrases pour 5 systèmes ou encore (HINTERLEITNER et al. 2011) avec deux groupes de 18 stimuli. Ce faible nombre de stimuli est généralement motivé par l'aspect particulièrement chronophage des campagnes d'évaluation perceptives. Quelques travaux récents ont mis en doute la méthodologie d'évaluation, comme LATORRE et al. 2014 qui étudie l'impact des références mentales des auditeurs sur les résultats des tests perceptifs, ou HINTERLEITNER et al. 2011 ; Mahesh VISWANATHAN et Madhubalan VISWANATHAN 2005 qui ont proposé des modifications des protocoles existants. Des alternatives aux méthodes classiques ont également été utilisées, sur le principe d'évaluations en ligne à grande échelle (*crowdsourcing*) comme décrit dans (BUCHHOLZ, LATORRE et YANAGISAWA 2013).

Un autre facteur que nous jugeons important, outre le nombre d'échantillons, est le fait de les choisir au hasard. En effet, choisir les échantillons au hasard revient à sélectionner uniquement les évènements les plus fréquents pour le système, et amène donc à une conclusion sur un comportement moyen. Si on compare deux systèmes assez proches, leur comportement moyen se trouve également très proche et les résultats n'apportent, en généralement, pas d'information car ils sont le plus souvent non significatifs.

1.2.2 Méthode d'évaluation

Partant de ce constat, pour révéler les différences entre deux systèmes, puisque c'est le but de l'évaluation, il faut se concentrer sur les différences constatées dans les signaux de parole générés. Lorsque les évaluations se fondent sur un petit ensemble d'échantillons, les signaux les plus différents sont bien souvent absents. C'est également le cas si les échantillons sont choisis au hasard. Par conséquent, nous proposons ce qui suit :

1. Synthétiser un grand nombre de textes provenant de domaines variés et de styles différents avec chaque système ;
2. Calculer pour chaque paire d'échantillons un coût d'alignement (par exemple une DTW – *Dynamic Time Warping* (Sakoe et Chiba 1978)) ;
3. Sélectionner les échantillons les plus dissemblables pour évaluer les systèmes.

Un point crucial de cette méthode est que la mesure de coût ne fait aucune hypothèse sur la qualité des systèmes. Elle ne fait que mesurer une différence entre deux signaux acoustiques. Cela permet de concentrer l'évaluation sur des zones d'intérêt, pour lesquelles les systèmes se comportent de manière différente.

Cette méthode a été utilisée avec succès pour la comparaison de systèmes de synthèse de type statistique (ici HTS) et de systèmes de type sélection d'unités (Chevelu et al. 2015 ; Chevelu et al. 2016). Dans ces expérimentations, un grand nombre de phrases de test a été généré (environ 27000). Plusieurs évaluations de type AB ont été conduites en choisissant des échantillons au hasard, en choisissant ceux avec le coût d'alignement le plus faible et ceux avec le coût d'alignement le plus fort. Le coût d'alignement a été calculé comme le coût d'une DTW entre les séquences de vecteurs MFCC normalisés par la longueur du chemin d'alignement. Les résultats montrent clairement qu'en choisissant de manière pertinente les échantillons, on peut séparer des systèmes assez proches et faire émerger une préférence.

De la même manière, cette méthode a été utilisée dans le cadre de la réduction de corpus (Chevelu et Lolive 2015). Les travaux de Lambert, Braunschweiler et Buchholz 2007 concluaient que la sélection aléatoire de phrases était aussi performante que l'application d'une stratégie de couverture des phénomènes linguistiques, remettant ainsi en cause l'intérêt de la réduction de corpus. Étant convaincus que l'absence de différence provenait de la méthodologie d'évaluation et du tirage aléatoire des échantillons, nous avons comparé :

- Un système de synthèse reposant sur un corpus pris dans son intégralité ;
- Un système de synthèse reposant sur un corpus réduit couvrant les diphonèmes ainsi qu'un ensemble d'attributs utilisés par le système de synthèse (phonème en fin de phrase, phonème en fin de groupe, phonème dans onset, phonème dans coda) ;
- Un système de synthèse reposant sur un corpus couvrant uniquement les diphonèmes et complété de manière aléatoire pour qu'il ait la même taille que le précédent.

Des évaluations de type MUSHRA ont été conduites avec ces trois systèmes. Les résultats montrent nettement que le choix éclairé des échantillons à évaluer permet de montrer que l'application de la réduction de corpus permet d'améliorer la qualité de la sortie du moteur de synthèse. Étant donné la proximité des systèmes de synthèse, ce résultat n'apparaît pas si on choisit les échantillons de manière aléatoire.

Cette méthode est maintenant utilisée dans la majorité de nos travaux et a également été utilisée avec d'autres critères que le coût d'alignement par DTW. Cependant, il est nécessaire de trouver un moyen généralisant la méthode à un nombre arbitraire de systèmes à comparer.

1.3 Conclusion

Dans ce chapitre, nous avons présenté les travaux menés sur le développement d'outils nécessaires pour le traitement de parole. La première contribution porte sur la librairie ROOTS qui répond à un problème de structuration des corpus afin de garantir la cohérence des données et leur utilisation facile pour la création de prototypes de recherche. Nous l'utilisons pour tous les travaux menés que ce soit pour la modélisation de la prononciation, l'analyse de la prosodie, la construction de voix de synthèse ou encore dans le moteur de synthèse. La deuxième contribution porte sur l'évaluation des systèmes de synthèse. La proposition permet de comparer des systèmes assez similaires, en se focalisant sur leurs différences. Les résultats obtenus montrent l'efficacité de la méthode.

Ces deux résultats offrent des perspectives intéressantes. La librairie apporte un gain en terme de représentation de l'information et en conséquence permet d'effectuer des requêtes précises et autorise l'analyse de grands corpus de manière efficace. La maturité de la librairie permet sa diffusion et son utilisation dans des contextes applicatifs assez variés dès lors qu'il est nécessaire de représenter des informations organisées en séquences. Des collègues de l'université de Saarland, Allemagne, utilisent ROOTS et en développent une version légère dans l'objectif de l'intégrer à MARYTTS. Les perspectives de développement de la librairie incluent l'ajout de la gestion de treillis et de graphes. Enfin, la méthodologie d'évaluation, sous l'hypothèse d'une généralisation à N systèmes, pourrait être appliquée dans le cadre du challenge *Blizzard*.

Chapitre 2

Modélisation de la prononciation

*Les travaux présentés dans ce chapitre ont été conduits principalement à l'IRISA dans le cadre de la thèse de Raheel Qader ainsi que du post-doctorat de Marie Tahon au sein du projet ANR SynPaFlex, dont je suis le coordinateur. Pour ces travaux, j'ai contribué sur le plan scientifique mais également sur le plan organisationnel en coordonnant les travaux. Les publications suivantes en sont le résultat : (*Qader *et al. 2014 ;* Qader *et al. 2015 ;* Qader *et al. 2016 ;* Tahon *et al. 2016a ;* Tahon *et al. 2016b). Les travaux suivants, réalisés en collaboration avec Gwénolé Lecorvé, ont permis de mener ces études (*Lecorvé *et* Lolive *2015 ;* Lecorvé *et* Lolive *2016).*

Le processus de synthèse nécessite d'être en mesure de prédire la séquence de phonèmes correspondant à la prononciation d'un texte donné en entrée. Pour réaliser cela, l'approche classique consiste à construire un dictionnaire de prononciations, qui recense la prononciation d'un sous-ensemble des mots de la langue, puis à élaborer des règles afin d'être en mesure de prédire la prononciation de mots hors du dictionnaire. Les règles peuvent bien entendu être construites manuellement ou apprises à l'aide de modèles statistiques.

Cependant, dans le contexte de la synthèse de parole, il est souhaitable de pouvoir générer une prononciation différente selon le style recherché ou encore le locuteur. En effet, la prononciation des mots est influencée par différents facteurs. Par exemple, il existe des différences notables entre une interaction dans un cadre formel ou non, une prise de parole préparée ou spontanée, ou encore entre un locuteur du sud de la France et un locuteur d'une autre région.

Dans ce chapitre, nous abordons les contributions dans le domaine de la modélisation de la prononciation ainsi que son adaptation en posant un cadre méthodologique rendant cela possible. Le cas de la parole spontanée est ensuite traité à travers une application sur un corpus de parole en anglais américain. Ensuite, la problématique de l'adaptation

à la voix d'un locuteur pour la synthèse est traitée.

2.1 Contexte

Les variantes de prononciation de mots ou énoncés ne sont pas prises en compte par les lexiques et modèles de prononciation utilisés dans les systèmes actuels de synthèse de la parole. Cela limite alors leur capacité à produire des signaux expressifs, notamment pour retranscrire un style spontané.

Les travaux connexes en production de variantes de prononciation peuvent être résumés d'après le type de l'approche retenue et la nature des informations utilisées. Tout d'abord, diverses approches par apprentissage automatique ont déjà été utilisées : des arbres de décision (Fosler-Lussier 1999 ; Vazirnezhad, Almasganj et Ahadi 2009), des forêts aléatoires (Dilts 2013), des réseaux de neurones (Chen et Hasegawa-Johnson 2004 ; Karanasou et al. 2013), des modèles de Markov cachés (Prahallad, Alan W Black et Mosur 2006) et des Champs Aléatoires Conditionnels (CAC) (Karanasou et al. 2013). Pour aller plus loin, d'autres travaux ont également proposé de combiner différentes techniques (Vazirnezhad, Almasganj et Ahadi 2009 ; Kolluru et al. 2014). Il est malheureusement difficile de comparer ces travaux car ceux-ci partagent rarement les mêmes données ou la même tâche.

Quant aux informations utilisées, des caractéristiques acoustiques peuvent être extraites à partir de signaux de parole d'un style visé et prises en compte pour l'adaptation de prononciations (fréquence fondamentale, énergie, durée, débit de parole...) (Bates et Ostendorf 2002 ; Bell, Brenier et al. 2009 ; Bell, Jurafsky et al. 2003), tandis que des informations linguistiques peuvent être dérivées de textes (distinction entre mots-outils et mots pleins, probabilité des mots, informations syllabiques, accentuation lexicale dans certaines langues...) (Vazirnezhad, Almasganj et Ahadi 2009 ; Bell, Brenier et al. 2009 ; Bell, Jurafsky et al. 2003). Récemment, Dilts 2013 a présenté une étude poussée sur la combinaison de ces deux types d'informations. Ce travail est proche du nôtre mais diffère dans le sens où la technique d'apprentissage automatique est différente et l'objectif principal était uniquement de réduire les prononciations. En complément, notons également que Chen et Hasegawa-Johnson 2004 ont montré que les informations d'un phonème canonique doivent être enrichies par celles de leur voisinage pour aboutir à de meilleures adaptations.

Enfin, il est important de noter que la plupart des travaux du domaine visent la reconnaissance automatique de la parole alors que les approches pour la synthèse sont encore rares et qu'aucune ne fait un usage des informations linguistiques aussi intensif que le nôtre. Pour la synthèse de parole, on peut citer (Brognaux et al. 2014) qui étudie l'impact des variations de prononciation sur la qualité de la synthèse pour différentes situations de communication. Ce travail montre un impact réel sur la qualité perçue du signal de synthèse.

2.2 Adaptation de la prononciation

Nous proposons de résoudre le problème de production de variantes de prononciation sous l'angle de l'adaptation d'une prononciation dite canonique [1] pour prendre en compte des éléments spécifiques comme le style (e.g. spontané), le locuteur, etc. L'objectif est d'obtenir une prononciation adaptée qui pourra ensuite être utilisée dans le moteur de synthèse.

La méthode proposée consiste à prédire pour chaque phonème d'une prononciation canonique si celui-ci doit être supprimé, remplacé, conservé ou complété par des phonèmes à insérer. Expérimentalement, la qualité d'une adaptation se mesure alors à son taux d'erreurs entre les phonèmes prédits, dits *adaptés*, et phonèmes effectivement réalisés dans le corpus.

Pour cela, cette méthode repose sur des Champs Aléatoires Conditionnels (CAC), dont l'utilisation est très répandue pour la phonétisation de mots ou d'énoncés (Wang et S. King 2011 ; Illina, Fohr et Jouvet 2011 ; Lecorvé et Lolive 2015). Ce type de modèle est particulièrement adéquat pour l'apprentissage sur des données séquentielles et symboliques et permet d'intégrer et combiner simplement un très large panel d'informations. Les différentes pistes qui ont été explorées pour cela sont illustrées par la figure 2.1. Nous les présentons ci-dessous et introduisons les questions qui s'y rattachent.

1. Chaque phonème p_i à prédire dépend de n caractéristiques $\{c_i^1, \dots, c_i^n\}$, par exemple le phonème canonique à adapter, sa position dans la syllabe ou la fréquence du mot qui le contient. La question est alors de savoir quelles caractéristiques parmi toutes celles considérées sont pertinentes et quelles autres dégradent l'adaptation.
2. L'ensemble des caractéristiques peut être étendu au voisinage de p_i, par exemple en considérant également le phonème canonique précédent et le suivant, ainsi que les caractéristiques qui leur sont associées. Cette notion de voisinage se définit en pratique par une fenêtre de taille W autour de p_i. Le choix de cette taille est un point que nous avons étudié.
3. Une dépendance entre p_i et la précédente prédiction p_{i-1} peut être considérée. Cela doit notamment permettre d'éviter des enchaînements de phonèmes possiblement non articulables. Nous avons cherché à savoir si cette dépendance est utile.
4. Enfin, il est possible d'interdire ou autoriser la propagation des dépendances par-delà les frontières de mots. Nous avons également étudié ces deux options.

Les CAC permettent naturellement de modéliser tous ces différents types de dépendances (Lafferty, McCallum et Pereira 2001) et nous avons pour cela utilisé l'outil Wapiti (Lavergne, Cappé et Yvon 2010).

1. Il s'agit de la prononciation produite par le phonétiseur indépendamment de tout style particulier.

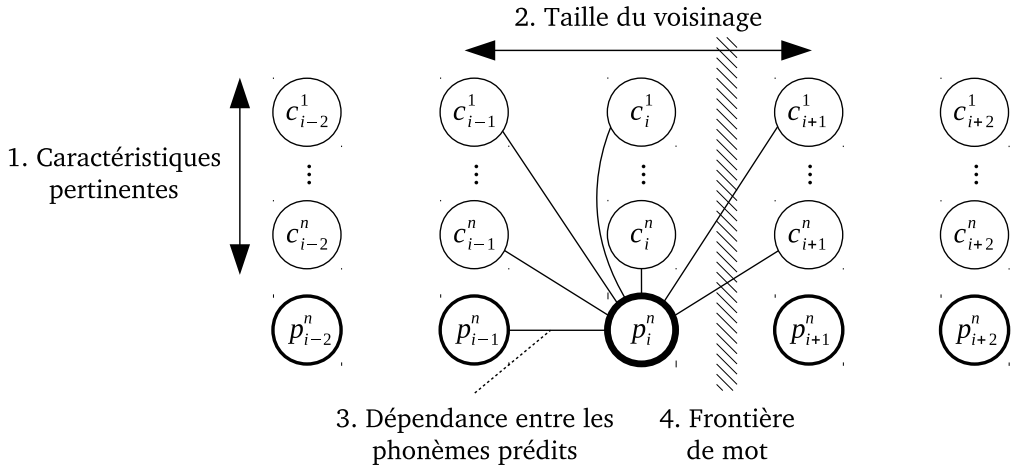

FIGURE 2.1 – *Vue d'ensemble des dépendances et paramètres à traiter pour l'apprentissage des CAC. Les nœuds et arrêtes représentent respectivement différentes informations et leurs liens de dépendance.*

2.3 Cas de la parole spontanée

Nous avons appliqué cette méthodologie pour le cas de la parole spontanée dans le cadre des travaux de thèse de Raheel Qader (QADER et al. 2014 ; QADER et al. 2015 ; QADER et al. 2016). Dans ce cadre, nous avons travaillé avec le corpus de parole conversationnelle Buckeye (PITT et al. 2005). La méthodologie suivie ainsi que les résultats sont reportés dans la suite de cette section.

2.3.1 Le corpus Buckeye

Ce corpus, en anglais, consiste en 40 entretiens non préparés avec des locuteurs de l'Ohio, aux États-Unis, chaque entretien durant 1 heure. 20 entretiens ont été sélectionnés au hasard, les autres ayant été laissés de côté pour d'éventuels futurs travaux. Les signaux de parole sont fournis avec des transcriptions vérifiées manuellement : une transcription orthographique et deux transcriptions phonétiques, l'une correspondant aux phonèmes canoniques qui auraient dû être prononcés si le style avait été neutre, l'autre aux phonèmes effectivement réalisés par le locuteur dans un cadre de parole spontanée. Chaque locuteur représente environ 7 400 mots et 22 800 phonèmes. Les phonèmes canoniques et réalisés ont été alignés automatiquement. Il en découle que 30 % des phonèmes et 57 % des mots sont prononcés différemment de ce qui était attendu.

Le corpus a été complété par des annotations automatiques, conduisant à l'ensemble des caractéristiques linguistiques, articulatoires et acoustiques dont le détail est donné

Attribut	Valeurs
a. Attributs linguistiques (23)	
phonème canonique	**40 phonèmes possibles**
position du phonème dans la syllable	**entier**
position inversée du ph. dans la syllabe	**entier**
accentuation lexicale de la syllabe	**aucun, primaire, secondaire**
partie de la syllabe	**onset, nucleus, coda**
position de la syllabe dans le mot	**initiale, interne, finale**
mot	**mot**
fréquence du mot en anglais	**fréquent, moyen, rare**
mot vide (d'après une liste) ?	**vrai, faux**
position du phonème dans le mot	**début, milieu, fin**
graphème	graphème
type de syllabe	ouverte, fermée
longueur du mot en syllabes	entier
fréquence du mot dans l'entretien	fréquent, moyen, rare
fréquence de la racine dans l'entretien	fréquent, moyen, rare
occurrence du mot dans l'entretien	entier
position du mot dans l'énoncé	entier
position inverse du mot dans l'énoncé	entier
longueur du mot en graphèmes	integer
fréquence de la racine en anglais	fréquent, moyen, rare
classe grammaticale	étiquette de classe grammaticale
position de l'énoncé dans l'entretien	entier
position inverse de l'énoncé dans l'entretien	entier
b. Attributs articulatoires (9)	
voyelle/consonne	**voyelle, consonne**
mode d'articulation	**nasale, plosive, fricative, etc.**
point d'articulation (consonnes)	**bilabiale, labiodentale, dentale, etc.**
point d'articulation (voyelles)	**antérieure, quasi-antérieure, etc.**
degré d'aperture	**fermé, pré-fermé, pré-ouvert, etc.**
voisement	**vrai, faux**
caractère arrondi ou non	**vrai, faux**
affriqué	vrai, faux
doublé	vrai, faux
c. Attributs acoustiques (10)	
énergie de la syllable	**fort, moyen, bas**
mouvement intonatif de la syllable	**montant, plat, descendant**
ratio pause/syllabe	**élevé, moyen, faible**
ton associé au phone	**t0, t1, t2, t3, t4, t5**
distance à la prochaine pause	**près, moyen, éloigné**
distance à la pause précédente	**près, moyen, éloigné**
distance à la prochaine hésitation	près, moyen, éloigné
distance à l'hésitation précédente	près, moyen, éloigné
ton de la syllabe	t0, t1, t2, t3, t4, t5
débit de parole	élevé, moyen, faible

TABLE 2.1 – *Liste des attributs linguistiques, articulatoires et acoustiques. Les attributs sélectionnés, ayant un nombre de votes* ≥ 20*, sont en gras.*

par la table 2.1. Afin d'être compatible avec l'emploi de CAC, les fréquences ont été catégorisées à masses de probabilité équivalentes en « fréquent », « moyen » et « rare ». Nous présentons maintenant la méthode en elle-même.

2.3.2 Choix et impact des paramètres

Sélection des attributs

L'entraînement de CAC à partir de trop nombreuses caractéristiques peut conduire à du sur-apprentissage. Par ailleurs, le temps de calcul et la quantité de mémoire nécessaires à l'entraînement est exponentiel en fonction du nombre de caractéristiques. Ainsi, nous avons effectué une sélection des attributs linguistiques à considérer. Cette sélection s'est faite en recherchant le meilleur ensemble de caractéristiques, c.-à-d. celui qui conduit au plus petit PER, pour chaque locuteur. Nous avons pour cela mis en place un mécanisme de vote. Une caractéristique a reçu un vote par nombre de fois où elle appartenait au meilleur ensemble d'un locuteur. Pour rendre ce processus robuste, deux stratégies de recherche ont été testées pour chaque locuteur. La première consiste à débuter avec un ensemble d'attributs réduit (uniquement l'étiquette du phonème canonique) et à ajouter les attributs un à un jusqu'à obtenir l'ensemble optimal. À chaque itération, tous les attributs sont évalués pour choisir et ajouter le plus performant. La seconde stratégie consiste en une élimination itérative des attributs inutiles en partant de l'ensemble complet. À l'issue des votes, il a arbitrairement été décidé de sélectionner les caractéristiques qui avaient reçu au moins 50 % des votes, soit 20 votes dans notre cas. La table 2.1 reporte en gras les caractéristiques qui ont finalement été sélectionnées ainsi, pour chaque groupe d'attributs. Pour les attributs linguistiques, il apparaît que les informations relatives aux syllabes et aux fréquences des mots sont les plus importantes. Ces conclusions sont cohérentes avec de précédents travaux (Adda-Decker et al. 2005 ; Vazirnezhad, Almasganj et Ahadi 2009 ; Bell, Brenier et al. 2009). Presque la totalité des caractéristiques articulatoires sont conservées. Enfin, les attributs acoustiques liés à l'intensité de la syllabe, au mouvement intonatif ainsi qu'à la position par rapport aux pauses sont conservés. On aurait pu s'attendre à ce que l'attribut lié au débit de parole soit conservé, mais ce n'est pas le cas ici. La table 2.2 compare les PER obtenus (i) sans adaptation et (ii) avec adaptation uniquement sur les phonèmes canoniques, (iii) des caractéristiques sélectionnées et (iv) de toutes les caractéristiques possibles, pour chaque groupe de caractéristiques. Il ressort clairement des résultats que la sélection des caractéristiques permet d'améliorer les résultats, excepté pour les attributs articulatoires, tout en réduisant le nombre d'attributs utilisés.

Effet du contexte

Pour la prédiction de la prononciation, l'intégration d'informations liées au contexte semble pertinent. Afin de le vérifier expérimentalement, nous avons défini le voisinage

Baseline (no adaptation)		28.3
Canonical phoneme only (with adaptation)		30.7 (+2.4)
+ Linguistic	Selected features (9)	25.1 (-3.2)
	All features (23)	26.6 (-1.7)
+ Articulatory	Selected features (7)	30.8 (+2.5)
	All features (9)	30.9 (+2.6)
+ Acoustic	Selected features (6)	26.7 (-1.6)
	All features (10)	27.1 (-1.2)

TABLE 2.2 – *PER (%) sur l'ensemble de développement sans adaptation ou avec adaptation et différents jeux de caractéristiques. Les tests sont conduits sur des mots isolés. Les différences absolues avec la séquence canonique sont reportées entre parenthèses.*

comme une fenêtre symétrique[2] de W phonèmes canoniques à gauche et à droite autour du phonème à adapter. Une fenêtre $W = 0$ signifie ainsi qu'aucun voisinage n'est considéré et $W = \pm 2$ que 5 phonèmes sont considérés au total (1 au centre, 2 à gauche et 2 à droite). La figure 2.2 présente les PER obtenus pour différentes valeurs de W. Ces résultats sont présentés soit dans le cas où les fenêtres ne peuvent pas traverser des frontières de mots (mots isolés), soit dans celui où elles le peuvent (énoncés). Les CAC ont été appris à partir des seuls phonèmes canoniques et dans la configuration unigramme (pas de dépendances entre phonèmes prédits). Il apparaît que la prise en compte du voisinage améliore significativement les résultats mais qu'un plateau est vite atteint lorsque W augmente. Ainsi, la valeur $W = \pm 2$ est retenue pour les expériences finales. Par ailleurs, contrairement à l'intuition, il semblerait que les adaptations mot à mot produisent de meilleurs résultats que lorsque l'adaptation se fait à l'échelle d'un énoncé entier. L'explication de ce phénomène est que les frontières de mots portent une information utile quant à la position d'un phonème canonique dans son mot. L'utilisation conjointe de fenêtres et d'énoncés supprime cette information.

Information inter-mots

En plus des variations de prononciation à l'intérieur des mots, il existe également des variations qui dépassent le cadre des mots, notamment en parole spontanée. Cela peut être observé lorsqu'un mot est entouré par certains mots spécifiques. Par exemple, le phonème /t/ dans le mot « *what* » (/wʌ**t**/) est parfois prononcé comme une occlusive glottale sourde lorsqu'il est suivi par le mot « *I* » (/aɪ/) comme dans « *what I'* » : /wʌ**ʔ** aɪ/.

En conséquence, dans cette étude, nous avons également modélisé la dépendance entre mots. Pour vérifier que cette information est utile, nous avons conduit des évaluations sur l'ensemble de développement en utilisant uniquement le phonème canonique. Les

2. Nous avons également testé des fenêtres dissymétriques mais celles-ci n'ont mené qu'à de moins bons résultats.

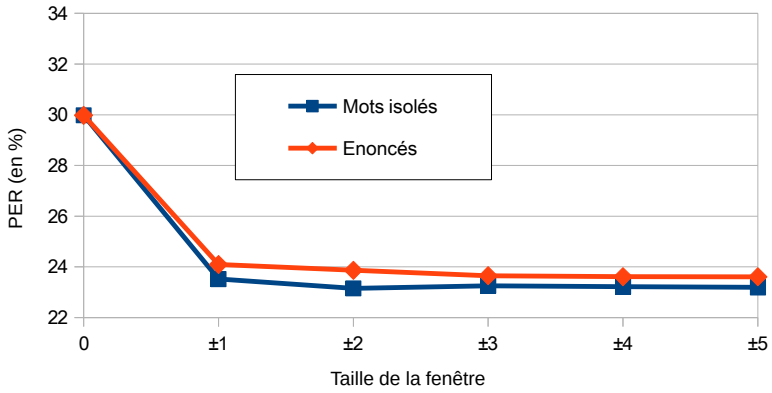

FIGURE 2.2 – *PER sur l'ensemble de développement en fonction de la taille de la fenêtre, pour les mots isolés et les phrases.*

résultats montrent une amélioration de 0.3 point lorsque cette information est incluse avec des PER de 30.7% et 30.4% respectivement pour les mots isolés et les énoncés.

Combinaison des paramètres

Dans ce paragraphe, une évaluation en combinant les paramètres précédents est effectuée sur les ensembles de test pour chaque locuteur du corpus. Ainsi, les modèles d'adaptation sont appris en combinant différentes configurations, à savoir :

- à partir des seuls phonèmes canoniques ou en incluant aussi les caractéristiques sélectionnées ;
- sans ou avec prise en compte du voisinage ($W = 0$ ou $W = \pm 2$) ;
- en considérant des successions de mots isolés ou, au contraire, des énoncés continus.

Les résultats obtenus avec ces différentes combinaisons sont présentés dans le tableau 2.3. Deux séries d'expériences ont été réalisées pour les mots isolés et les énoncés. Tout d'abord, les groupes d'attributs sélectionnés sont évalués séparément. Dans un second temps, toutes les combinaisons sont évaluées. Pour les mots isolés et les énoncés, les PER obtenus sont comparés avec la séquence de référence c'est-à-dire sans adaptation.

On peut observer, à la fois sur les mots isolés et sur les énoncés, que l'adaptation avec des attributs linguistiques permet une légère amélioration par rapport à l'adaptation avec uniquement le phonème canonique (*ligne 2, Tableau 2.3*). La différence sur les énoncés est plus grande en raison de l'absence de l'information sur la frontière de mot (*ligne 1*). Concernant les attributs articulatoires, on peut noter que leur ajout (*ligne 3*) dégrade les résultats sur les mots isolés et n'apporte pas d'amélioration sur les énoncés par rapport à l'adaptation avec le phonème canonique seulement. Les caractéristiques acoustiques (*line 4*) amènent une nette amélioration, avec une réduction de 3.1 points sur les mots isolés et 4.4 points sur les énoncés par rapport à l'adaptation avec le phonème canonique seulement. Bien que leur extraction, effectuée directement à partir du signal,

					Mots isolés	Énoncés
	Phonème canonique (sans adaptation)				28.3	28.0
	Ph. canonique	Ling.	Artic.	Acous.		
1	✓				24.2 (-4.1)	25.2 (-2.8)
2	✓	✓			24.0 (-4.3)	23.7 (-4.3)
3	✓		✓		24.4 (-3.9)	25.2 (-2.8)
4	✓			✓	21.5 (-6.8)	22.0 (-6.0)
5	✓	✓	✓		24.0 (-4.3)	24.1 (-3.9)
6	✓	✓		✓	**21.1 (-7.2)**	**20.8 (-7.2)**
7	✓		✓	✓	21.4 (-6.9)	22.0 (-6.0)
8	✓	✓	✓	✓	21.2 (-7.1)	21.1 (-6.9)

TABLE 2.3 – *PERs (%) sur l'ensemble de test pour toutes les combinaisons de jeux d'attributs sur des mots isolés et des énoncés. Les différences absolues avec la séquence canonique sont reportées entre parenthèses.*

puisse avoir un rôle dans ce résultat, cela montre leur importance pour l'adaptation de la prononciation.

La seconde série d'expériences combine les différents groupes d'attributs et montre que le groupe d'attributs articulatoires, même combiné avec d'autres informations (*lignes 5, 7*), n'apporte pas de meilleurs résultats. Au contraire, les attributs linguistiques et acoustiques (*lignes 5, 6, 7*) améliorent toujours les résultats. Que ce soit pour les mots isolés ou les énoncés, la combinaison des attributs linguistiques et acoustiques (*ligne 6*) donne les meilleurs résultats avec 21.1% et 20.8% sur les mots isolés et les énoncés respectivement.

De manière globale, il ressort que :

- les attributs acoustiques ont la plus grande influence sur l'adaptation de la prononciation,
- les attributs articulatoires n'apportent pas d'information additionnelle par rapport au phonème canonique,
- bien que les informations linguistiques apportent un gain limité, elles permettent d'obtenir un gain supplémentaire lorsqu'elles sont combinées à d'autres attributs,
- l'information inter-mot apporte un gain limité (en particulier avec les attributs linguistiques et acoustiques) mais significatif[3].

En outre, la prononciation est souvent variable pour un locuteur. En conséquence, il ne semble pas suffisant d'évaluer un modèle par rapport à une unique séquence de phonèmes de référence. Si on s'intéresse au n premières hypothèses produites par les CAC, on peut mesurer la qualité de la meilleure prédiction parmi cet ensemble d'hypothèses et obtenir

3. Les p-values sont 6.889×10^{-4} et 8.005×10^{-4} en utilisant un test t à échantillons jumelés et un test de Wilcoxon, respectivement, avec un niveau de confiance $\alpha = 0.05$.

n ▶	1	2	3	4	5	10	20	30	40	50
Ph. canonique	30.4	23.0	19.0	16.3	14.8	10.5	7.8	6.6	6.0	5.6
+ cara. ling.	24.3	17.1	13.8	11.8	10.4	7.3	5.3	4.5	4.0	3.6
+ cont.	23.8	16.5	13.3	11.2	9.9	6.9	4.8	4.0	3.5	3.2
+ cara. ling + cont.	23.6	16.4	13.1	11.1	9.8	6.8	4.9	4.1	3.7	3.3

TABLE 2.4 – *PERs oracle pour les n-meilleures hypothèses avec des mots isolés, pour n entre* 1 *et* 50 *(en %).*

ce que l'on peut qualifier de PER oracle. L'idée est d'évaluer le nombre d'hypothèses nécessaires pour s'approcher de la prononciation cible. Le tableau 2.4 présente l'évolution des PER oracle pour un nombre d'hypothèses allant de 1 à 50.

Le premier constat est que l'ajout de seulement deux hypothèses permet d'atteindre un PER d'environ 16% et l'ajout de 5 hypothèses permet de descendre en dessous des 10% d'erreur. Cela semble montrer que dans beaucoup de cas, le modèle parvient à prédire une prononciation tout à fait possible parmi les premières hypothèses. Cela milite pour l'introduction d'un modèle de réordonnancement. Cependant, même si cela peut améliorer les résultats comme nous allons le voir par la suite, cela n'est pas suffisant pour évaluer la qualité d'une prononciation. En effet, la variabilité du phénomène implique que l'utilisation d'une référence unique n'est pas suffisante. De nouvelles façons de mesurer l'erreur ou la qualité d'une prononciation semblent donc nécessaires.

Dépendance avec le phonème prédit

Les CAC peuvent prendre en compte des dépendances entre phonèmes en utilisant des configurations de type bigramme et uni+bigramme. Le tableau 2.5 présente la comparaison de ces deux configurations avec la configuration unigramme pour les mots isolés et les énoncés. On peut noter à partir des résultats que la configuration bigramme provoque une augmentation du PER pour toutes les expériences. En particulier, dans les cas des attributs linguistiques+acoustiques où le PER augmente de 11.1 points pour la configuration bigramme et de 1.3 point pour la configuration uni+bigramme. Ce comportement est sans doute dû à la rareté de certains bigrammes de phonèmes dans l'ensemble d'apprentissage. De plus, les PER augmentent encore plus lorsque le nombre d'attributs utilisés lors de l'apprentissage augmente. Ces résultats montrent qu'il est préférable, dans notre cas, de ne pas utiliser ce type de dépendances avec les CAC.

Réordonnancement des hypothèses

Pour vérifier cette explication et approfondir l'intérêt de tenir compte des dépendances entre phonèmes prédits, nous avons conduit une autre série d'expériences. Nous avons appris un modèle de langage (ML) sur les phonèmes réalisés de l'ensemble des données

Phonème canonique (sans adaptation)			28.3
	Phonème canonique	Linguistique	Linguistiue+acoustique
Unigramme	24.5 (-3.8)	24.0 (-4.3)	21.5 (-6.8)
Bigramme	30.9 (2.6)	32.3 (4.0)	32.6 (4.3)
Uni+bigramme	24.8 (-3.5)	24.6 (-3.7)	22.8 (-5.5)

TABLE 2.5 – *PERs (%) sur l'ensemble de développement pour les configurations unigramme, bigramme et uni+bigramme sur les configurations avec phonème canonique, attributs linguistiques et attributs linguistiques+acoustiques. Les différences absolues avec la séquence canonique sont reportées entre parenthèses.*

d'apprentissage [4], puis avons utilisé ce modèle pour réordonner *a posteriori* les meilleures hypothèses d'adaptation fournies par nos CAC. Précisément, chaque hypothèse h est associée à un score $s(h)$ calculé comme une interpolation logarithmique des probabilités fournies par le CAC et par le ML, comme suit :

$$s(h) = \Pr_{\text{CAC}}(h) \times \Pr_{\text{ML}}(h)^{\alpha} \times \beta^{n} \ , \tag{2.1}$$

où α et β sont deux paramètres à optimiser et n est le nombre de phonèmes dans h. Le facteur β sert à contrebalancer le favoritisme naturel du ML envers les hypothèses les plus courtes. Le réordonnancement consiste alors à sélectionner l'hypothèse de score s le plus élevé. En pratique, le ML est un modèle 5-gramme avec un lissage de Witten-Bell et α et β ont été optimisés de sorte à minimiser le PER sur l'ensemble de développement. L'apprentissage du ML, l'optimisation des paramètres et le réordonnancement ont été effectués grâce à l'outil SRILM (STOLCKE et al. 2011). La table 2.6 présente les PER obtenus avant et après réordonnancement sur l'ensemble de test. Les CAC utilisés sont ceux qui produisent les meilleurs résultats à la table 2.3. Ceux-ci n'incluent aucune dépendance entre phonèmes prédits. Il apparaît clairement que l'introduction de ces dépendances par le ML permet d'obtenir des améliorations significatives du PER. Ces résultats confortent en outre notre hypothèse sur l'effet négatif d'un trop grand nombre de paramètres lors de l'apprentissage des CAC.

2.3.3 Évaluation subjective de l'adaptation

Afin d'évaluer l'impact de cette approche sur des échantillons de parole synthétique, une évaluation perceptive a été conduite avec 10 locuteurs anglais natifs. L'objectif du test est d'essayer de répondre aux questions suivantes : Quels sont les degrés de *spontanéité* et d'*intelligibilité* perçus dans les échantillons synthétisés en comparant des échantillons générés avec la prononciation canonique, la prononciation réelle et la prononciation générée par adaption avec différentes configurations. Plus précisément, les

4. Un seul ML pour les 20 locuteurs a été appris plutôt qu'un par locuteur afin d'estimer des probabilités fiables.

Mots isolés

Phonème canonique (sans adaptation)	28.3	
	Avant réord.	Après réord.
Phonème canonique	24.2 (-4.1)	23.7 (-4.6)
+ Attributs linguistiques	24.0 (-4.3)	23.7 (-4.6)
+ Attributs linguistiques + acoustiques	21.1 (-7.2)	**20.6 (-7.7)**

Énoncés

Phonème canonique (sans adaptation)	28.0	
Phonème canonique	25.2 (-2.8)	25.0 (-3.0)
+ Attributs linguistiques	23.7 (-4.3)	23.5 (-4.5)
+ Attributs linguistiques + Acoustiques	20.8 (-7.2)	**20.7 (-7.3)**

TABLE 2.6 – *PERs (%) pour les configurations avec le phonème canonique, les attributs linguistiques, les attributs linguistiques+acoustiques, avant et après réordonnancement (réord.) en utilisant les* 10 *meilleures hypothèses sur l'ensemble de test. Les différences absolues avec la séquence canonique sont reportées entre parenthèses.*

configurations testées reposent soit seulement sur le phonème canonique, soit en ajoutant les caractéristiques linguistiques, soit encore en ajoutant également les attributs acoustiques. Toutes les configurations incluent le réordonnancement.

L'évaluation repose sur des tests de type AB contenant chacun 40 étapes avec 2 étapes introductives additionnelles permettant au testeur de se familiariser avec la tâche à réaliser. À chaque étape, deux échantillons de parole sont présentés et deux questions sont posées. Le testeur doit, en premier lieu, décider lequel des deux échantillons est le plus spontané, et ensuite lequel est prononcé de la manière la plus intelligible. Pour ces deux questions, les testeurs peuvent aussi indiquer qu'ils n'entendent pas de différence entre les échantillons. Le texte des échantillons prononcé est également affiché aux testeurs à chaque étape afin de rendre la compréhension du contenu plus simple.

Les énoncés ont été sélectionnés parmi les 2000 énoncés de l'ensemble de test de manière à ce qu'ils contiennent des différences importantes entre la prononciation canonique et la prononciation réalisée. Ce choix permet d'assurer que les phrases sélectionnées sont celles pour lesquelles les locuteurs parlent de manière très spontanée. Cette manière de choisir n'introduit pas de biais dans les tests perceptifs car aucun système adapté n'intervient dans cette sélection. De plus, nous avons montré dans (CHEVELU et al. 2015) que cette méthode de sélection permet d'être plus efficace que la sélection aléatoire d'échantillons, cette dernière pouvant résulter dans la comparaison d'échantillons très similaires, voire identiques. Enfin, afin d'aider les testeurs à se concentrer sur la prononciation, les phrases très longues ou très courtes ont été filtrées, et le texte des phrases a été contrôlé afin d'éviter des phrases incorrectes grammaticalement. Les phrases sélectionnées ont été synthétisées grâce à un système reposant sur HTS, version 2.2, avec l'ensemble d'attributs standards, entraîné sur le corpus de parole du challenge Blizzard de 2012 (Simon KING et KARAISKOS 2012). L'utilisation de ce corpus de parole, issu de livres audios, permet

d'avoir un corpus avec une variabilité raisonnable sans posséder les difficultés liées à la parole spontanée.

La figure 2.3 montre la comparaison des échantillons de parole générés par le système sans adaptation contre les systèmes adaptés et réalisé en termes de spontanéité et d'intelligibilité. L'intelligibilité est ici évaluée d'une manière comparative entre deux échantillons, en donnant le texte au testeur. Il s'agit d'une manière non classique d'évaluer ce critère mais justifiée ici par la difficulté de la tâche dans le cas de parole spontanée. Sur la figure 2.3, chaque configuration *testée* contre le système sans adaptation est présentée par une barre verticale dont les segments dénotent le pourcentage de préférence pour la configuration testée, le système de référence ou aucun des deux. Tout d'abord, on peut noter que les prononciations réalisées sont logiquement jugées comme plus spontanées que la référence mais bien moins intelligibles. Concernant les prononciations adaptées, la configuration utilisant seulement le phonème canonique ne donne pas de bons résultats que ce soit pour la spontanéité ou l'intelligibilité. À l'opposé, les deux autres configurations adaptées sont jugées plus spontanées que le système de référence, mais provoquent cependant une dégradation de l'intelligibilité. Enfin, l'adaptation semble avoir d'aussi bons résultats avec ou sans les attributs acoustiques. Cette conclusion est intéressante puisque dans un contexte de synthèse, l'intégration d'attributs acoustiques fiables prédits à partir du texte semble assez irréaliste.

De manière complémentaire, la figure 2.4 compare les prononciations réalisées aux prononciations générées par adaptation. Les résultats comparés à la prononciation canonique sont également reportés sur cette figure (de manière identique à la figure 2.3). De manière surprenante, les configurations adaptées avec les attributs linguistiques et linguistiques+acoustiques semble être comparables aux prononciations réalisées en terme de spontanéité. Ce résultat montre que l'adaptation permet de refléter un style spontané, sans avoir forcément une prononciation identique à la prononciation réalisée. De plus, les échantillons résultant des prononciations réalisées sont toujours évalués comme moins intelligibles. Cette conclusion supporte l'idée selon laquelle plus la parole est spontanée, moins elle est intelligible. Enfin, on peut encore une fois noter que la configuration reposant sur les attributs linguistiques uniquement apporte de meilleurs résultats, en particulier pour l'intelligibilité.

Pour finir, les résultats des tests perceptifs permettent de conclure que l'approche proposée, et ses différentes étapes (sélection des attributs, combinaison des attributs, réordonnancement des hypothèses) semblent être efficaces afin de modifier la prononciation pour qu'elle reflète un style spontané. Malgré tout, la méthode peut encore être améliorée afin de préserver l'intelligibilité, cette fois-ci, afin d'améliorer la qualité de la parole en sortie de la synthèse, ce qui n'était pas l'objectif dans ces travaux. Ainsi, une piste pour améliorer la qualité de la synthèse est de tenir compte du contenu de la voix utilisée pour synthétiser les échantillons. En effet, certaines séquences de phonèmes très spontanées peuvent être très éloignées des habitudes du locuteur, donc du corpus utilisé pour créer la voix de synthèse et ainsi conduire à une baisse importante de la qualité de sortie.

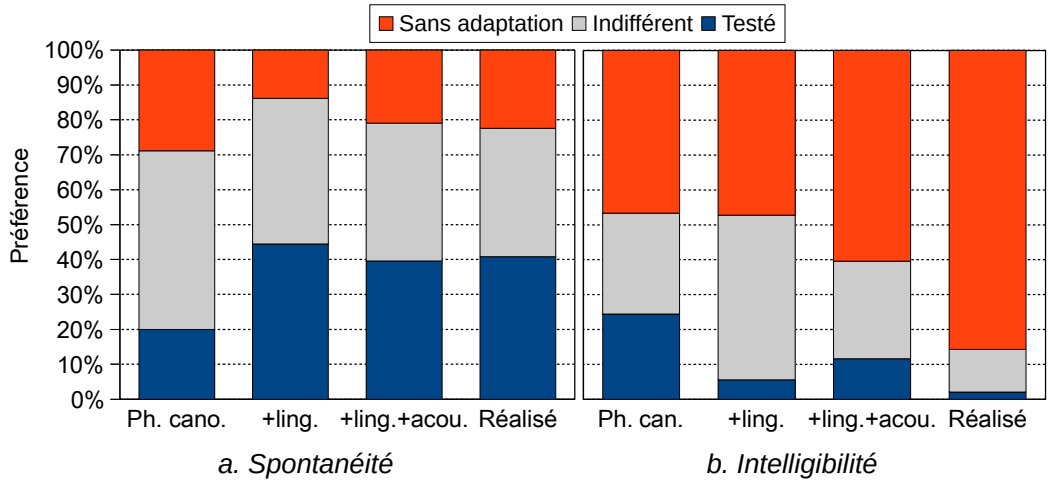

FIGURE 2.3 – *Comparaisons entre la prononciation canonique (sans adaptation) et les prononciations adaptées ou réalisées en termes de spontanéité et d'intelligibilité. Trois configurations adaptées sont utilisées : phonème canonique seul (Ph. cano.), phonème canonique + attributs linguistiques (+ling.), phonème canonique + attributs linguistiques + attributs acoustiques(+ling.+acou.).*

2.4 Adaptation au corpus de synthèse

Une condition essentielle pour obtenir une parole de qualité est la cohérence entre la chaîne phonétique prédite par le phonétiseur et le contenu du corpus utilisé pour construire la voix. Pour obtenir cette cohérence, on peut « forcer » le locuteur à avoir la « bonne » prononciation, c'est-à-dire celle prédite par le phonétiseur, ou bien adapter la sortie du phonétiseur aux habitudes du locuteur. C'est cette deuxième approche que nous explorons ici en appliquant la méthodologie présentée précédemment.

Dans ce cadre, on considère le problème de l'adaptation de la sortie du phonétiseur au style contenu dans la voix utilisée pour la synthèse. Cette problématique est commune aux techniques de synthèse par concaténation et également statistiques.

Dans cette section, nous détaillons des expérimentations menées sur ce sujet dans le cadre du projet ANR SynPaFlex (voir B.2), dont l'objet est l'amélioration de la flexibilité du processus de synthèse de parole.

2.4.1 Méthodologie générale

L'objectif de cette étude est donc de réduire les différences entre les phonèmes produits en sortie du phonétiseur, dénotés *phonèmes canoniques*, et les phonèmes contenus dans le corpus de parole, dénotés *phonèmes réalisés*. Pour cela, la méthodologie présentée précédemment, publiée dans (QADER et al. 2015 ; QADER et al. 2016), est ici adaptée au

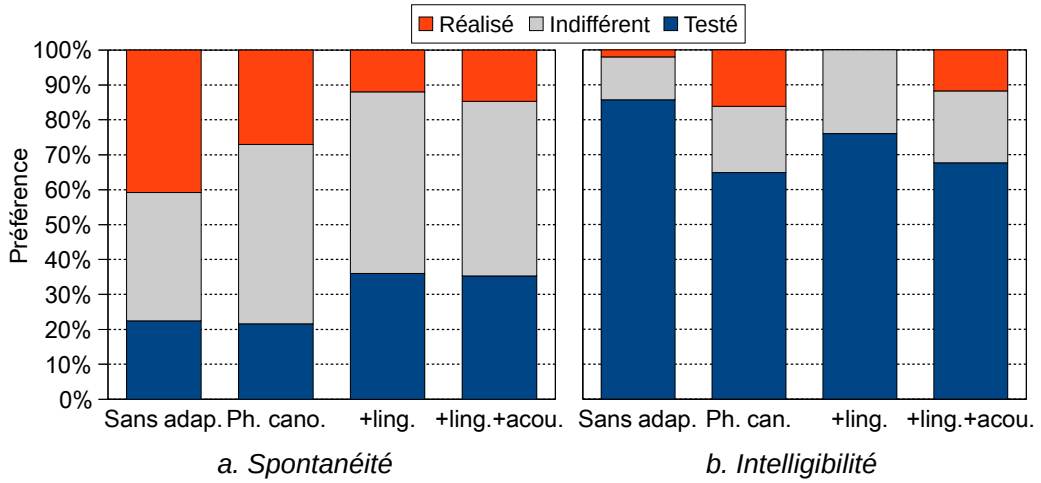

FIGURE 2.4 – *Comparaisons entre la prononciation réalisée et les prononciations adaptées ou canonique (Sans adap.) en termes de spontanéité et d'intelligibilité. Trois configurations adaptées sont utilisées : phonème canonique seul (Ph. cano.), phonème canonique + attributs linguistiques (+ling.), phonème canonique + attributs linguistiques + attributs acoustiques (+ling.+acou.).*

Français. Un modèle *CAC* est entraîné de manière à prédire une séquence de phonèmes adaptée au corpus à partir des phonèmes canoniques. Pour cela, un corpus de parole est utilisé et divisé en deux parties : 70% pour l'apprentissage et 30% pour la validation. L'ensemble d'apprentissage est de plus sub-divisé en 7 plis utilisés pour sélectionner et combiner les attributs dans des conditions de validation croisée. Chaque modèle est appris sur 6 plis, le septième est utilisé pour tester le modèle. L'ensemble de validation est quant à lui conservé en vue de l'évaluation finale ainsi que des tests perceptifs.

Un ensemble de 52 attributs de nature linguistique, phonologique, articulatoire et prosodique est sélectionné en appliquant un algorithme de sélection incrémental sous condition de validation croisée. La totalité des attributs et la méthode de sélection sont présentés dans (TAHON et al. 2016a). Ces attributs sont ensuite combinés pour obtenir le jeu d'attributs final.

Corpus de parole

Les expériences sont menées sur un corpus de parole, *IVS*, en langue française conçu pour les systèmes vocaux interactifs. Ainsi, ce corpus couvre tous les diphonèmes présents en français ainsi que les mots les plus usités dans le domaine des télécommunications. La voix est celle d'une femme et peut être qualifiée de « neutre ». L'échantillonnage de la voix est de 16kHz et sa taille est d'environ 6h40'. Il comprend 7208 phrases, 225080 instances de phonèmes. La prononciation a été fortement contrôlée pendant l'enregistrement. La

Groupe d'attributs	**# attr.**	**Attributs sélectionnés**
Linguistique (L)	2	Mot ♦ Radical
Phonologique (Ph)	7	Syllabe canonique ♦ Position de la syllabe dans le mot ♦ Position inverse du phonème dans la syllabe (numérique) ♦ Position du phonème (avant et inverse) (numérique) ♦ Longueur du mot en phonèmes (numérique) ♦ Ratio pauses/syllabes (faible, normal, élevé)
Articulatoire (A)	0	-
Prosodique (Pr)	6	Énergie de la syllabe (faible, normal, élevé) ♦ Ton de la syllabe et du phonème (de 1 à 5) ♦ contour de F_0 du phonème (descendant, plat, montant) ♦ Débit de parole (faible, normal, élevé) ♦ Distance à la pause précédente (de 1 à 3)

TABLE 2.7 – *Attributs sélectionnés dans le cas du français avec le corpus IVS.*

segmentation en phonèmes du corpus a été réalisée automatiquement et corrigée manuellement.

Attributs sélectionnés

Les attributs sélectionnés sont présentés dans le tableau 2.7. Au total, 15 attributs linguistiques, phonologiques et prosodiques sont sélectionnés. Tout d'abord, seuls 2 attributs linguistiques ont été sélectionnés : le mot et le radical. Comme ces deux attributs sont fortement corrélés, nous aurions pu nous attendre à ce que seulement l'un des deux soit sélectionné. Cependant, comme le mentionne (GUYON et ELISSEEFF 2003), la sélection d'attributs redondants peut contribuer à réduire le bruit de classification et ainsi renforcer la séparation des classes. Les autres attributs liés au mot, comme sa fréquence d'apparition en français, n'ont reçu que très peu de votes. De manière surprenante, il apparaît qu'aucun attribut articulatoire n'a été sélectionné. Un nombre plus important de votes pour ces attributs pouvait être attendu en raison des études, notamment récentes, montrant l'intérêt de ces attributs pour la modélisation des variations de prononciation (LIVESCU, JYOTHI et FOSLER-LUSSIER 2016). Un ensemble de 7 attributs phonologiques est inclu dans l'ensemble final. La plupart d'entre-eux sont liés à la position du phonème dans la phrase. Aucune caractéristique liée à la syllabe (comme la position dans la syllabe, sa structure ou son type) n'a été sélectionnée.

Finalement, 6 attributs prosodiques sur les 7 attributs évalués ont été sélectionnés. Dans un système réel, la prosodie devrait également être prédite à partir du texte. Cependant,

Phonème canonique (sans adaptation)		11.5 [0.0]
Phonème canonique seul (C)		6.9 [-4.6]
Linguistique (C+L)	tous (18)	4.4 [-7.1]
	selectionnés (2)	4.4 [-7.1]
Phonologique (C+Ph)	tous (17)	4.5 [-7.0]
	selectionnés (7)	4.6 [-6.9]
Articulatoire (C+A)	tous (9)	7.1 [-4.4]
	selectionnés (0)	-
Prosodique (C+Pr)	tous (7)	4.8 [-6.7]
	selectionnés (6)	4.8 [-6.7]
C + L + Ph	selectionnés (9)	4.0 [-7.5]
C + L + Pr	selectionnés (8)	3.5 [-8.0]
C + Ph + Pr	selectionnés (13)	3.6 [-7.9]
C + L + Ph + Pr	selectionnés (15)	3.2 [-8.3]

TABLE 2.8 – *PER moyen obtenus sur les 7 plis de l'ensemble d'apprentissage. L'évolution en points de pourcentage par rapport à la prononciation canonique est donnée entre crochets.*

comme cette tâche est encore un problème de recherche, les attributs prosodiques sont extraits directement à partir du signal de parole. Procéder de cette manière permet de vérifier à quel point la prosodie affecte les modèles de prononciation. Dans le cas présent, ce résultat est cohérent avec l'état de l'art et suggère qu'un modèle prosodique performant est utile pour la prédiction de la prononciation d'un locuteur.

Les PER moyens sur les 7 plis du corpus d'apprentissage sont reportés dans le tableau 2.8. L'utilisation d'un modèle de prononciation entraîné avec le phonème canonique seul permet d'améliorer de $4,6$ points le PER et de montrer que l'adaptation permet de réduire l'inconsistance entre la sortie du phonétiseur et le contenu du corpus de parole. L'ajout des groupes d'attributs de manière séparée permet de réduire encore le PER, à l'exception des attributs articulatoires. De manière intéressante, la réduction du nombre d'attributs dans chaque groupe n'affecte pas le taux d'erreur. La réduction la plus intéressante de PER apparaît avec les attributs linguistiques. Seulement 2 attributs permettent de réduire le PER de $7,1$ points comparé à la prononciation canonique.

La combinaison des groupes d'attributs permet d'améliorer les résultats. La combinaison des attributs prosodiques et linguistiques sélectionnés apporte une réduction significative du PER de $8,0$ points avec seulement 8 attributs. La combinaison des trois groupes donne les meilleurs résultats avec une diminution du PER de $8,3$ points. Au final, seulement un tiers des attributs de l'ensemble complet est conservé.

Phonème canonique (sans adaptation)		11.2 [0.0]
Phonème canonique seul (C)		6.6 [-4.6]
C + L + Ph	selectionnés (9)	3.9 [-7.3]
C + L + Ph + Pr	selectionnés (15)	3.3 [-7.9]

TABLE 2.9 – *PER obtenus sur l'ensemble de validation. L'évolution en points de pourcentage par rapport à la prononciation canonique est donnée entre crochets.*

Évaluation de l'adaptation sur les données de validation

Les résultats obtenus sont sensiblement les mêmes que dans la section précédente. La même configuration est celle qui inclut les attributs prosodiques (- 7,9 pp). Cependant, la différence apportée par la prosodie (0, 4pp) n'est pas aussi importante que lors de la phase d'apprentissage (0, 8pp). De plus, la robustesse des attributs linguistiques et phonologiques est certainement meilleure sur des données non vues que les attributs prosodiques. Considérant ces résultats, on peut s'attendre à ce que le modèle incluant les attributs prosodiques améliore le plus la qualité de la parole synthétique.

La plupart des confusions entre la prononciation canonique et les phonèmes réalisés concerne des allophones : o ≒ ɔ, e ≒ ɛ et ɛ̃ ≒ œ̃. Ces confusions ne peuvent pas être considérées comme des erreurs en français en raison de leur apparition possible suivant le style de parole. De manière similaire, une part importante des insertions concerne le schwa, qui est connu pour sa possible élision. D'autres substitutions concernent des choix d'étiquetage ou encore d'alphabet, par exemple ɲ ≒ nj, ə ≒ ø. Les suppressions concernent principalement les liaisons, comme t, z qui sont peu prédits par le phonétiseur alors qu'ils sont prononcés systématiquement dans le corpus de parole. L'utilisation des modèles de prononciation permet de réduire de telles confusions de manière automatique.

2.4.2 Évaluation subjective de l'impact de l'adaptation à la voix

Afin d'évaluer la qualité des échantillons de parole générés avec la prononciation adaptée, une évaluation perceptive a été conduite avec 14 locuteurs français natifs. L'évaluation repose sur des tests de type AB avec 40 échantillons pour lesquels chaque testeur doit répondre à la question suivante : « *entre A et B, lequel des deux échantillons est de meilleure qualité ?* ». Les réponses possibles sont *A*, *B*, or *pas de préférence*. Les phrases utilisées pour les tests sont sélectionnées de manière aléatoire en sous-échantillonnant l'ensemble de validation en suivant la distribution du PER entre les prononciations canoniques et réalisées. Les échantillons de parole ont été synthétisés d'une part avec le système par sélection d'unités décrit dans (GUENNEC et LOLIVE 2014a) et d'autre part en utilisant HTS v2.2 avec des attributs standards (ZEN et al. 2007). Cinq prononciations sont évaluées : phonèmes canoniques sans adaptation, phonèmes adaptés avec les phonèmes canoniques (C), adaptation en ajoutant les attributs linguistiques et phonolo-

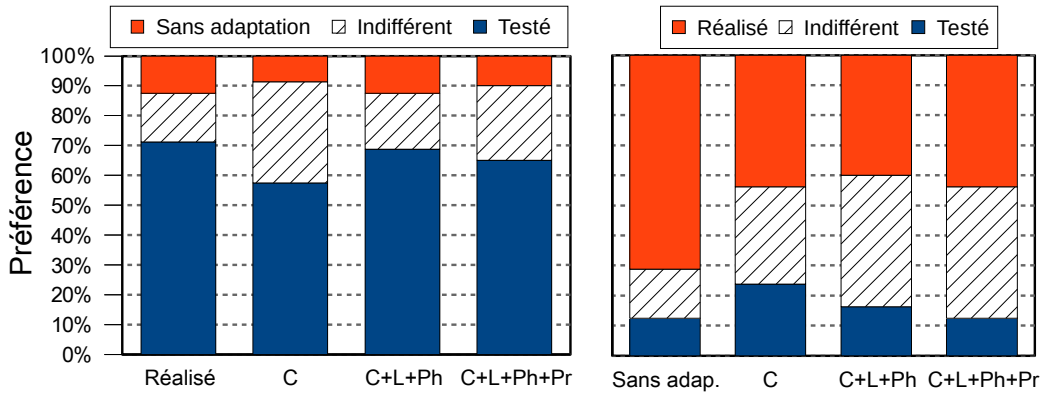

FIGURE 2.5 – *Résultats des tests AB pour la synthèse par sélection d'unités. À gauche : proportion de préférence pour le système testé {réalisé, C, C+L+Ph, C+L+Ph+Pr} comparé à la prononciation canonique. À droite : proportion de préférence pour le système testé {sans adaptation, C, C+L+Ph, C+L+Ph+Pr} comparé à la prononciation réalisée.*

giques (C+L+Ph), adaptation en ajoutant les attributs prosodiques (C+L+Ph+Pr) et les phonèmes réalisés tels qu'annotés dans le corpus de parole.

Les figures 2.5 et 2.6 présentent les résultats, respectivement, pour la sélection d'unités et la synthèse paramétrique. Pour chaque figure, la partie gauche présente la comparaison entre les systèmes testés (réalisé, C, C+L+Ph, C+L+Ph+Pr) contre la prononciation canonique, tandis que la figure de droite compare les systèmes testés (sans adaptation, C, C+L+Ph, C+L+Ph+Pr) à la prononciation réalisée.

Sur les deux figures, nous pouvons observer que les prononciations adaptées et réalisées sont largement préférées par rapport à la prononciation canonique. Cela montre clairement que les modifications effectuées permettent d'améliorer la qualité de la synthèse en augmentant l'adéquation avec le contenu du corpus de parole. Par ailleurs, les résultats montrent que les prononciations adaptées, sans être équivalentes à la prononciation réalisée, permettent de réduire fortement la préférence pour la prononciation réalisée. Cela confirme que dans de nombreux cas, l'adaptation permet de produire une prononciation équivalente à celle réalisée effectivement et d'atteindre le même niveau de qualité. De plus, les attributs prosodiques, qui contribuent à une amélioration du PER, ne permettent pas d'améliorer la qualité de la parole générée dans les évaluations conduites. Dans la mesure où la prédiction des attributs prosodiques est difficile, ce résultat est intéressant puisqu'il permet de relativiser leur importance pour la prédiction de la prononciation.

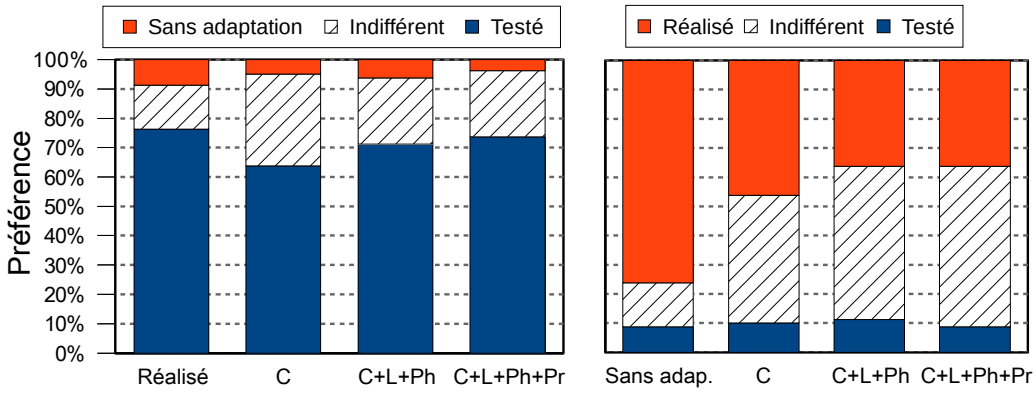

FIGURE 2.6 – *Résultats des tests AB pour la synthèse avec HTS. À gauche : proportion de préférence pour le système testé {réalisé, C, C+L+Ph, C+L+Ph+Pr} comparé à la prononciation canonique. À droite : proportion de préférence pour le système testé {sans adaptation, C, C+L+Ph, C+L+Ph+Pr} comparé à la prononciation réalisée.*

2.4.3 Étude de l'impact de la quantité de données sur l'adaptation

En raison du coût élevé que représentent les corpus annotés manuellement, il est utile d'évaluer quelle est la quantité minimale de données d'apprentissage nécessaires pour réaliser l'adaptation de la prononciation, avec la méthodologie proposée. Les résultats devraient apporter de l'information sur la précision attendue en fonction de la taille du corpus d'entraînement. Pour une taille de corpus donnée, les modèles sont évalués en termes de PER sous condition de validation croisée.

2.4.4 Protocole

L'ensemble d'apprentissage du corpus précédant, qui contient 70% du corpus initial, est découpé en $N_f = 7$ plis : 6 sont utilisés pour l'apprentissage et 1 pour le développement. Les différentes tailles de corpus évaluées sont obtenues en découpant le corpus d'entraînement original en 2 fois 7 sous-ensembles, puis 4 fois 7 sous-ensembles, 8 fois 7 sous-ensembles, etc. À chaque étape, 6 plis sont utilisés pour l'entraînement des modèles, le septième est conservé pour le développement. Afin de limiter les temps de calcul, nous avons limité N_f à 100 pour des durées d'ensemble d'apprentissage inférieures à 300 minutes. Lorsque la quantité de données d'apprentissage diminue, N_f augmente rendant ainsi les résultats plus fiables : de 243.3 min. de données d'apprentissage ($N_f = 7$, 4321 phrases) à 40 s. de données d'apprentissage ($N_f = 100$, 12 phrases). L'ensemble de validation est quant à lui composé de 120.2 min. de données, soit 2161 phrases.

Deux ensembles d'attributs (C et CLPrPh) ainsi que deux tailles de fenêtre (W0 et W2) sont évalués. L'objectif est ainsi d'évaluer l'effet du nombre d'attributs sur la précision et estimer le risque que présente l'usage d'un nombre important d'attributs sur un ensemble

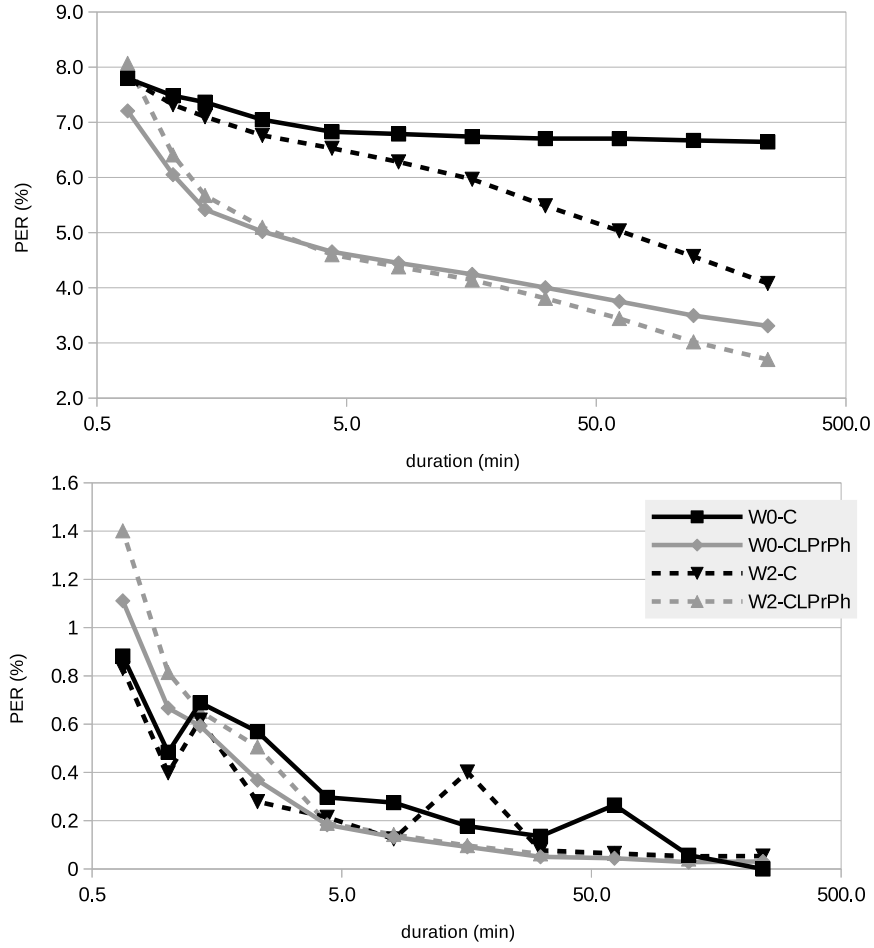

FIGURE 2.7 – *Moyenne (en haut) et écart-type (en bas) du PER entre la prononciation canonique et adaptée obtenus sur l'ensemble de validation. La moyenne est calculée pour tous les plis disponibles pour une durée d'ensemble d'apprentissage donnée (échelle logarithmique). Pour mémoire, la prononciation canonique présente une erreur de* 11.2 *%.*

d'apprentissage trop restreint.

2.4.5 Résultats

Comme attendu, les résultats moyens présentés sur la figure 2.7 montrent que plus la durée de l'ensemble d'entraînement est réduite, plus le PER est élevé. De manière surprenante, des modèles entraînés avec très peu de données permettent de diminuer le PER de 4.0 points par rapport au modèle de référence, qui correspond à la meilleure configuration (W0-CLPrPh). Ainsi, les ensembles d'apprentissage restreints permettent de corriger les erreurs les plus courantes : prononciation récurrente, changement d'al-

Taille appr.	Rég. Lin.	W0-C	W0-CLPrPh	W2-C	W2-CLPrPh
> 0.7 min	Slope	-0.17	-0.54	-0.58	-0.73
	Corr. coef.	0.74	0.85	0.99	0.86
> 4.0 min	Slope	-0.04	-0.34	-0.62	-0.48
	Corr. coef.	0.96	1.00	0.99	0.99

TABLE 2.10 – *Régression linéaire entre le PER et la durée de l'ensemble d'apprentissage (échelle logarithmique).*

phabet, schwa et liaison du français. De plus, il a été constaté qu'avec des ensembles trop petits, les résultats varient beaucoup en raison d'un contenu qui ne représente pas suffisamment de phénomènes. En conséquence, il est possible d'apprendre les règles les plus fréquentes avec des corpus de taille réduite mais il faut veiller à ce que leur contenu soit suffisamment représentatif des phénomènes observés.

L'évolution du PER en fonction de la taille du corpus d'apprentissage a un comportement différent suivant que la durée du corpus est supérieur ou non à 4.4 min. De manière intéressante, pour une durée de corpus d'entrainement supérieure à ce seuil, l'évolution de la courbe logarithmique du PER est quasi-linéaire (coefficient de corrélation > 0.96, cf. tableau 2.10). Ce résultat est en accord avec ceux obtenus dans le cadre d'expériences en reconnaissance de la parole (MOORE 2003).

Pour des durées faibles (< 4.4 min.), l'usage d'une fenêtre sur les phonèmes n'a quasiment aucun effet sur le PER, tandis que le nombre d'attributs en a un. Lors de l'apprentissage de modèles avec très peu de données, les effets de la fenêtre et de l'ensemble d'attributs sont limités (intervalle de variation du PER de 0.9 points). Ces résultats montrent que les modèles CAC, entraînés avec des jeux de données réduits, permettent tout de même une amélioration par rapport à la référence. Pour des ensembles d'apprentissage plus grands (> 4.4 min.), l'ensemble d'attributs utilisé a moins d'impact que la fenêtre sur les phonèmes. L'accroissement de la quantité de données n'améliore pas de manière significative les résultats pour les modèles entraînés avec la configuration sans fenêtre W0-C. Dans ce cas, multiplier par 10 la quantité de données améliore le PER de seulement 0.5 point. Pour les autres configurations, l'augmentation de la quantité de données provoque une amélioration significative. Une explication possible est que le nombre d'attributs ou la taille de la fenêtre augmentant, un accroissement de la taille de l'ensemble d'apprentissage est nécessaire afin de faire varier les différentes valeurs d'attributs.

Cependant, les résultats obtenus montrent qu'à partir d'environ 5 min. de corpus d'apprentissage, l'ajout de nouvelles données est coûteux et ne provoque pas d'amélioration importante. En effet, comme le PER est représenté pour une durée sur une échelle log-linéaire, un PER idéal de 0 serait obtenu avec un ensemble d'apprentissage de $3 \cdot 10^8$ heures de corpus en considérant la configuration W2-CLPrPh.

2.5 Conclusion

Dans ce chapitre, une méthodologie d'adaptation de la prononciation a été présentée, ainsi que son application à la modélisation des variantes de prononciation dans la parole spontanée, et l'adaptation de la prononciation en fonction d'un corpus de parole pour la synthèse.

Les outils utilisés, i.e. les CAC, sont des outils statistiques permettant de traiter des problèmes de prédiction de séquences d'étiquettes. En particulier, l'utilisation des CAC s'est déjà révélée être performante pour la prédiction de la prononciation. Leur utilisation ici permet de conserver une cohérence méthodologique entre conversion graphèmes vers phonèmes d'une part, et la prédiction de variantes de prononciation d'autre part.

La méthodologie présentée repose sur des attributs linguistiques, phonologiques, articulatoires et prosodiques. L'utilité des différents attributs et groupes d'attributs a été évaluée à la fois à travers des mesures objectives et lors de campagnes de tests perceptifs. Les résultats montrent l'utilité des attributs linguistiques et prosodiques. Néanmoins, dans les expériences menées, l'utilisation d'attributs prosodiques, même si elle apporte de meilleurs résultats sur des critères objectifs, n'a pas permis d'améliorer la perception de la parole synthétique.

De manière générale, pour la première application, nous avons montré qu'il est possible de modifier la prononciation afin de la rendre plus spontanée. Cependant, on se heurte rapidement à un problème d'intelligibilité et de qualité du signal de parole généré, nuisant à la perception de spontanéité. Des travaux sont nécessaires notamment sur la constitution de corpus adaptés à la parole spontanée, ce qui pose des problèmes de définition des conditions d'enregistrement pour garantir la spontanéité et également de la quantité de données disponibles pour chaque locuteur.

Pour l'adaptation de la prononciation à la voix de synthèse, les résultats semblent prometteurs. En effet, ils permettent d'améliorer la qualité de la parole synthétique de manière perceptible. Ce mécanisme d'adaptation automatique de la prononciation permet d'envisager une réduction des contraintes lors de la phase d'enregistrement ou de collecte des corpus de parole et donc en ayant un contrôle moindre sur la constance de la prononciation du locuteur. Ceci étant, une part de la complexité se reporte sur la phase de segmentation automatique qui doit pouvoir prendre en compte des variantes de prononciation de manière fiable.

En terme de perspectives, l'adaptation de la prononciation à des contenus expressifs en vue de la synthèse nécessite l'établissement d'un compromis entre degré d'expressivité de la prononciation générée et contenu du corpus de parole. À ce sujet, nous sommes à l'heure actuelle en train de chercher des solutions pour combiner un modèle expressif avec un modèle adapté à la voix. De plus, la technologie utilisée ici permet de générer simplement N hypothèses de prononciation. Une piste intéressante est l'intégration d'un treillis de prononciation de manière explicite lors de la synthèse afin de tenir compte

dynamiquement, non seulement des habitudes du locuteur, mais également de la qualité acoustique des unités présentes.

Chapitre 3

Styles de parole pour la synthèse

Les travaux présentés dans ce chapitre sont le fruit d'une collaboration avec Elisabeth Delais-Roussarie (DR, CNRS), Hiyon Yoo (Laboratoire de Linguistique Formelle - LLF) et Mathieu Avanzi (alors Post-doc au LLF) ainsi que des projets ANR Phorevox, centré sur l'apprentissage de l'écrit, et SynPaflex qui se focalise sur l'amélioration de l'expressivité en synthèse de parole. Notamment, la collaboration citée précédemment s'est focalisée sur l'adaptation des contraintes prosodiques pour la synthèse de parole (Mathieu Avanzi, Christodoulides *et al. 2014), sur l'étude du rythme à la fois en parole naturelle et en synthèse de parole (*Yoo, *Elisabeth* Delais-Roussarie *et al. 2015 ; Elisabeth* Delais-Roussarie, Lolive, Yoo *et* Guennec *2016a) ainsi que la construction automatique de dictées (Elisabeth* Delais-Roussarie, Lolive, Yoo, Barbot *et al. 2014 ;* Le Maguer *et al. 2014b ;* Le Maguer *et al. 2014a ;* Yoo, Le Maguer *et al. 2014).*

La prise en compte des styles de parole pour la synthèse de parole à partir du texte est une nécessité afin la rendre applicable dans de nombreux domaines et d'en améliorer l'expressivité. Plusieurs approches peuvent être utilisées afin d'aller vers cette prise en compte en synthèse. D'un côté, on trouve des approches statistiques comme présentées dans (Obin 2011) et de l'autre, des analyses expertes sur des ensembles de données réduits comme (Goldman, Auchlin et Simon 2009). Dans les deux cas, l'idée sous-jacente est de dériver des règles, ou des modèles, permettant de prédire au mieux les paramètres prosodiques qui doivent être utilisés.

Dans ce chapitre, nous commençons par présenter une analyse de quatre styles de parole en vue de la dérivation de règles simples applicables aisément dans un moteur de synthèse de parole par concaténation ou paramétrique. Dans un deuxième temps, une comparaison entre le rythme dans la parole naturelle et celui généré par un moteur de synthèse de parole est réalisée. Cette étude a pour but d'essayer de comprendre quelles sont les

Style de parole	Nb. locuteurs	Nb. syll.	Nb. tokens	Durée (sec.)
Contes (TAL)	6F/2H	5942	4189	1065.25
Dictée (DIC)	2F	4175	2918	893.56
Politique (POL)	3F/3H	6875	4539	1362.02
Roman (NOV)	2F/2H	7496	5226	1286.97
Total	13F/7H	24488	16872	4607.81

TABLE 3.1 – *Composition du corpus.*

différences en parole synthétique et parole naturelle et ainsi proposer des améliorations. Enfin, une application concrète de l'adaptation de règles est présentée dans le cadre de la génération automatique de dictées.

3.1 Comparaison de différents style de parole

Dans cette partie, une comparaison de quatre styles de parole, adressés à des enfants (dictée, contes) et à des adultes (romans, discours politiques) est effectuée. L'objectif est de faire apparaître, si possible, les grandes différences entre ces styles en utilisant un jeu de paramètres prosodiques limité. Ces paramètres sont choisis en fonction de deux critères : leur capacité à discriminer les styles de parole en Français, et la possibilité de les contrôler au cours du processus de synthèse.

3.1.1 Constitution du corpus

Contenu du corpus

Cette étude se focalise sur quatre styles de parole pour lesquels il s'agit de parole lue adressée à une audience donnée : lecture de contes (TAL), dictées (DIC), discours politiques (POL) et lecture de romans (NOV). Ces quatre styles diffèrent par l'audience, enfants ou adultes, et également par l'impact souhaité : importance d'être compris pour les dictées et de convaincre pour les discours politiques. 30 minutes de parole ont été collectées et analysées pour chaque style de parole. Le tableau 3.1 détaille le nombre de locuteurs et la durée exacte des échantillons du corpus. Tous les participants parlent un français « standard ».

Annotation et analyse des données

La segmentation en phones des signaux acoustiques a été vérifiée et corrigée manuellement. La transcription orthographique a ensuite été annotée en POS (Part-Of-Speech tags) afin d'assigner un statut phonologique à chaque mot, indiquant ainsi s'il peut être

accentué ou non de manière à segmenter ultérieurement les données en mots phonologiques (PW). De plus, les syllabes proéminentes ont été identifiées de manière automatique par l'algorithme Analor (Mathieu Avanzi, Obin et al. 2011) qui repose sur un ensemble réduit de paramètres acoustiques. Une annotation manuelle a également été réalisée et confrontée à l'annotation automatique pour décider du caractère proéminent ou non des syllabes. Finalement, les frontières des groupes accentuels (AP) et des groupes intonatifs (IP) ont été identifiées automatiquement comme décrit dans (Mathieu Avanzi, Christodoulides et al. 2014).

Pour cette étude, des paramètres prosodiques connus pour jouer un rôle significatif dans la différentiation des styles de parole en Français ont été extraits :

- Longueur des groupes accentuels (nombre de syllabes par AP) ;
- Longueur des groupes intonatifs (nombre de syllabes par IP) ;
- Ratio de syllabes proéminentes en début de groupe prosodique (PW) par rapport au nombre total de syllabes initiales de PW. De telles syllabes ont souvent été décrites comme caractéristiques du style didactique, voir par exemple (Fónagy 1980 ; Albert Di Cristo 1999) ;
- Taux d'articulation, calculé comme la durée moyenne syllabique par IP ;
- Durée des pauses et leur distribution ;
- Registre, calculé comme la différence entre le F_0 maximum et le F_0 minimum par IP, exprimé en demi-tons (ST).

3.1.2 Accentuation et phrasé

Le tableau 3.2 présente les longueurs moyennes des AP, IP ainsi que le ratio moyen de syllabes proéminentes en début de groupe prosodique. On peut noter que la longueur des AP est significativement moins importante pour le style *dictée* que pour le style *roman* ($p < .001$), mais cela n'est pas le cas pour les styles *conte* et *discours politique*. Aucune différence significative n'est trouvée entre les styles *discours politique*, *roman* et *conte*.

De plus, on peut noter que la longueur des IP dépend du style de parole. Ainsi, le style *dictée* diffère significativement du style *roman* ($p < .001$) et *conte* ($p < .01$), mais pas du style *discours politique*. Parallèlement, ce dernier diffère significativement uniquement du style *roman* ($p < .001$). Enfin, une différence significative est observée entre *roman* et *conte* ($p < .05$). De manière qualitative, le style *roman* présente des IP plus longs que les trois autres styles tandis que les styles *discours politique* et *conte* possèdent des longueurs d'IP similaires. Le style *dictée* possède des IP plus courts que les styles *roman* et *conte*.

Enfin, il semble qu'une dépendance existe également entre le style de parole et le ratio de syllabes proéminentes en début de PW. On observe que le style *dictée* présente un taux significativement plus élevé de syllabes proéminentes en début de PW que les trois

	DIC	POL	TAL	NOV
Longueur AP (syl/AP)	2.96 (1.25)	3.17 (1.5)	3.16 (1.28)	3.29 (1.34)
Longueur IP (syl/IP)	4.98 (3.05)	5.66 (3.77)	6.17 (3.68)	7.63 (4.7)
Ratio syl. proéminentes (%)	63.76 (2.01)	32.05 (1.52)	21.65 (1.41)	22.02 (1.31)

TABLE 3.2 – *Moyennes et écart type pour les longueurs d'AP, d'IP et pour les ratio de syllabes proéminentes en début de groupe prosodique. Résultats présentés pour les quatre styles de parole.*

autres styles ($p < .001$), qui ne présentent pas de différences entre-eux. Cela est cohérent avec les résultats concernant la présence d'un accent didactique en position initiale de mot ou de groupe.

Ces premiers résultats montrent que ces trois critères semblent permettre de différencier les styles de parole étudiés.

3.1.3 Taux d'articulation et pauses

Pour les variables temporelles, nous avons évalué le lien entre le style de parole et le taux d'articulation ainsi que la durée des pauses silencieuses et leur distribution. Les taux d'articulation pour chaque style au niveau de l'IP sont présentés sur la figure 3.1. Les styles *dictée* et *discours politique* sont très similaires du point de vue du taux d'articulation, avec des taux moyens respectifs de 182.36 ms/syl et 193.96 ms/syl. De la même manière, les styles *roman* et *conte* sont assez similaires avec des taux d'articulation moyens respectifs de 182.36 ms/syl et 193.96 ms/syl. Les styles *dictée* et *discours politique* possèdent une longueur moyenne des syllabes plus grande que pour *roman* et *conte*, ce qui tend à indiquer que les locuteurs articulent plus lentement pour ces deux styles.

Les longueurs des pauses silencieuses ont été modélisées par des mélanges de lois log-normales afin de voir si la répartition et les longueurs des pauses varient entre les styles de parole étudiés. Le nombre de composantes a été estimé pour chaque style grâce à un critère BIC. Les paramètres des modèles obtenus sont reportés dans le tableau 3.3. Les résultats font apparaître deux comportements différents pour ce qui est de la gestion des pauses. Pour les styles *conte* et *roman*, les pauses se répartissent sur deux modes : le premier représente des pauses courtes ($\approx 100ms$) peu fréquentes et le deuxième des pauses longues ($\approx 550 - 600$ms). Les deux autres styles, à savoir *dictée* et *discours politique*, présentent quant à eux trois modes distincts. Il est intéressant de noter la répartition équitable des durées des pauses sur chacun des modes pour le style *discours politique*. En particulier, les pauses longues du style *dictée* correspondent à des pauses inhérentes au style dictée permettant d'écrire, alors que pour le style *discours politique*, il s'agit plutôt de pauses liées au style rhétorique.

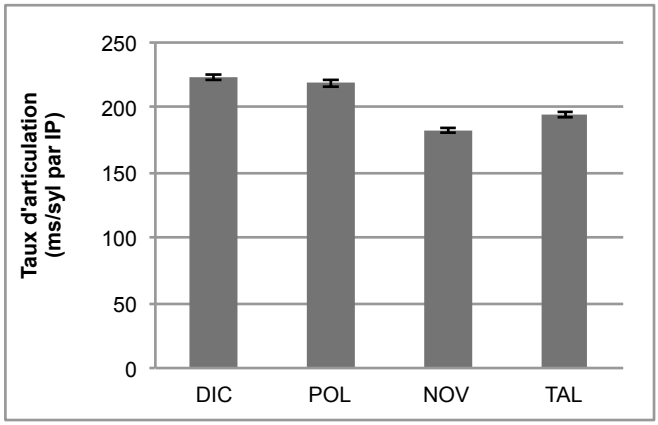

FIGURE 3.1 – *Taux d'articulation par IP (en ms/syl) en fonction du style de parole (DIC, POL, NOV and TAL).*

	DIC			POL			TAL		NOV	
α	13%	35%	52%	30%	39%	31%	9%	91%	12%	88%
μ (log ms)	2.04	2.54	2.96	2.18	2.62	3.01	1.95	2.74	2.03	2.78
σ (log ms)	0.11	0.15	0.18	0.16	0.14	0.13	0.18	0.28	0.16	0.26

TABLE 3.3 – *Paramètres de modèles à mélange de lois log-normales pour les styles de parole. Le nombre de composantes est estimé grâce au critère BIC.*

3.1.4 Registre

L'étude du registre moyen entre les différents styles permet de faire apparaître des différences significatives. En particulier, les styles *dictée* et *conte* possèdent un registre bien plus grand (8.33 et 7.71 demi-tons) que *discours politique* et *roman* (6.34 et 5.36 demi-tons, resp.). Une analyse statistique révèle de plus que le style *dictée* diffère significativement des styles *discours politique* ($p < .01$) et *roman* ($p < 0.01$). Une différence significative apparaît également entre les styles *conte* et *roman* ($p < .01$), ce qui n'est pas le cas entre *roman* et *discours politique*. Une étude complémentaire indique que le genre n'affecte pas le registre dans les données que nous avons étudiées.

3.1.5 Discussion

Les résultats obtenus pour l'accentuation et le phrasé indiquent que les locuteurs, pour les styles *dictée* et *discours politique*, ont une forte tendance à segmenter leur flux de parole en unités de taille plus réduite que pour les styles *roman* et *conte*. Les longueurs d'AP et d'IP obtenues sont en accord avec les études précédentes, et confirment que pendant une dictée ou un discours politique, les locuteurs tendent à aligner la longueur des IP sur celle des AP : en moyenne, un IP est 1.5 fois plus long qu'un AP pour ces styles,

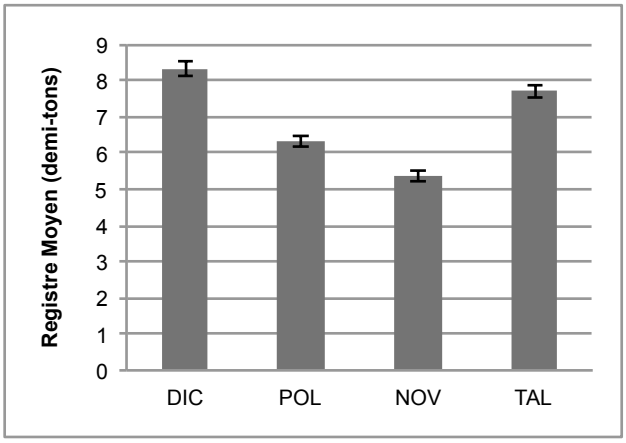

FIGURE 3.2 – *Registre moyen par IP (en semi-tons) en fonction du style de parole (DIC, POL, NOV and TAL).*

tandis que le ratio est de 2 pour les styles *roman* et *conte*. Néanmoins, un comportement différent est observable entre les styles *dictée* et *discours politique* pour ce qui du taux de syllabes proéminentes en début de groupe prosodique. Ce qui n'est pas le cas pour le style *discours politique* vis-à-vis de ce paramètre et des deux autres styles.

Pour les variables temporelles, des similitudes existent entre les styles *dictée* et *discours politique* et également entre les styles *roman* et *conte*. En effet, les locuteurs utilisent clairement trois niveaux de pauses pour les premiers, et seulement deux niveaux pour les derniers. Cela est confirmé par le fait que les styles *dictée* et *discours politique* présentent un taux d'articulation plus bas que les styles *roman* et *conte*. La baisse du taux d'articulation traduit ainsi le fait que les locuteurs parlent plus lentement, ce qui est cohérent avec ces deux styles.

Finalement, le registre est semblable entre les styles *dictée* et *conte* d'un côté, et aussi *discours politique* et *roman* de l'autre. Ces résultats indiquent que la parole adressée à des enfants tend à comporter une échelle de variation du pitch plus grande que de la parole adressée à des adultes.

En conséquence, si on considère un système de synthèse de parole standard, produisant un style de parole similaire au *roman*, qui peut être considéré comme le plus proche de la parole lue, les règles suivantes pourraient être utilisées pour modifier le modèle prosodique afin de produire de la parole ayant des propriétés similaires aux styles étudiés :

- *Dictée* : des AP et IP plus courts doivent être prédits, avec un taux de syllabes initiales de groupes prosodiques proéminentes bien plus important, un accroissement du nombre et de la longueur des pauses, et enfin un registre plus important. Ces règles semblent cohérentes avec le fait qu'il s'agisse d'un style didactique en favorisant le temps d'assimilation et les variations prosodiques permettant d'insister sur

la structure du texte.

- *Discours politique* : la longueur des IP doit être réduit de manière similaire au style *dictée*, et le taux de syllabes initiales de groupes prosodiques proéminentes augmenté, mais pas autant que pour le style *dictée*. De plus, ce style comporte également des pauses longues (autour d'une seconde) qui doivent être ajoutées.
- *Conte* : pour ce style, le paramètre le plus important est le registre qui doit être plus important que pour les autres styles, avec un taux d'articulation légèrement plus élevé. Les autres paramètres se comportent de la même manière que pour le style *roman*.

Un prolongement de ce travail, pour lequelle la méthodologie employée ici est transposable, est l'étude des émotions. Par exemple, (Bartkova, Jouvet et Elisabeth Delais-Roussarie 2016) étudie les différences entre les six émotions du *Big Six* et tente d'en fournir une caractérisation prosodique.

Comme nous le verrons en partie dans le chapitre 4, ces règles peuvent être implantées dans un moteur de synthèse de parole, que ce soit par concaténation ou paramétrique. Notamment, une intervention est nécessaire dans les modules de prédiction prosodique afin d'intégrer les consignes adéquates. Pour un système par concaténation, une intégration de ces éléments lors du processus de sélection des unités, notamment par la prise en compte de coûts cibles spécifiques, est nécessaire.

3.2 Patrons rythmiques et littéraires

Ces dernières décennies, la qualité globale de la parole synthétisée s'est améliorée de façon notable avec l'émergence de nouvelles techniques de synthèse comme la synthèse par corpus (A. J. Hunt et Alan W. Black 1996). Néanmoins, générer une prosodie naturelle qui tient compte des genres et styles de parole reste un challenge (Marc Schröder 2009 ; Obin 2011), en particulier pour les aspects rythmiques. De fait, la composante rythmique semble souvent peu naturelle en synthèse de parole et doit être améliorée pour permettre une meilleure utilisation de la synthèse dans de nombreuses applications (jeu vidéo, logiciel éducatif, lecture de livres audio, etc.).

Dans le projet de recherche Phorevox, visant à utiliser la synthèse de parole pour favoriser l'apprentissage de l'écriture à des enfants de cycle 2 (CP, CE1), il fallait améliorer le système de synthèse de parole afin qu'il puisse lire de façon claire et naturelle des contes, des poèmes et des comptines. Pour tenter d'atteindre cet objectif, nous avons comparé les patrons rythmiques obtenus en parole naturelle et en parole synthétique pour chacun des genres littéraires visés (comptines, poèmes, contes). Nous avions émis au départ l'hypothèse que les patrons rythmiques les plus précis seraient observés pour les contes, les corpus utilisés pour extraire les unités de parole lors de la synthèse contenant essentiellement des textes lus comparables à des récits. Cependant, les résultats obtenus

Genre littéraire	Nombre de mots	Nombre de syll.	Nombre de syll. effectif
Comptines	158	228	1137 (454 pour synt.)
Poèmes	290	422	2155 (808 pour synt.)
Contes	522	777	3861 (1538 pour synt.)
Total	970	1427	7153 (2800)

TABLE 3.4 – *Composition du corpus*

n'ont pas confirmé cette hypothèse, les lectures de comptines étant souvent plus satisfaisantes. Aussi, avons-nous tenté de comprendre pourquoi les patrons rythmiques sont plus adéquats dans le cas des poèmes et des comptines que dans le cas des contes, alors que les corpus utilisés pour générer les stimuli ne contenaient pas ce genre littéraire.

3.2.1 Corpus

Le corpus utilisé pour étudier les patrons rythmiques en parole naturelle et synthétique est constitué de trois types distincts de textes adressés à des enfants : six comptines, quatre poèmes et deux extraits de contes. Le tableau 3.4 présente la composition quantitative du corpus par genre littéraire. Les différences de réalisation entre les locuteurs (qui sont indirectement notées par l'écart entre le nombre de syllabes par genre à multiplier par cinq, c'est-à-dire le nombre de locuteurs, et le nombre effectif de syllabes) résultent principalement de l'insertion ou de l'élision de schwas, ou de l'omission d'un mot. Le nombre effectif de syllabes obtenu pour la parole synthétique est donné entre parenthèses.

L'ensemble des textes a été produit par cinq voix différentes (deux voix synthétiques et trois voix naturelles). Pour les voix naturelles, le corpus a été enregistré par trois locuteurs (deux hommes et une femme) dans un studio d'enregistrement. Les participants ont eu le temps de lire les textes et de les répéter avant l'enregistrement. Parmi ces locuteurs, deux ont lu les textes de la même manière que des parents liraient une histoire à leurs enfants, tandis que le troisième est un acteur confirmé et les a lus avec beaucoup plus d'expressivité. Les stimuli synthétisés ont été produits grâce au système de synthèse par corpus présenté dans (GUENNEC et LOLIVE 2014a). Deux voix de synthèse ont été utilisées pour cette étude :

- la voix d'homme SY-P, construite à partir de 10 heures de parole extraites d'un livre audio (un roman lu par un acteur) ;
- la voix de femme SY-A, construite à partir de 7 heures de parole lue, les éléments lus ayant été sélectionnés spécifiquement pour la construction d'un système de synthèse. Les différences de contenu et de taille des corpus amènent à considérer SY-P comme une voix plus expressive que SY-A, qui est plus neutre.

Pour générer les stimuli de synthèse, la structure des strophes et des vers dans les poèmes

et comptines a été représentée par des marques de ponctuation. Ainsi, pour obtenir la version synthétisée, les trois strophes sous (1), extraites du poème "La fourmi" de R. Desnos ont été mises en forme comme indiqué sous (2).

(1) *Une fourmi traînant un char*
plein de pingouins et de canards
ça n'existe pas, ça n'existe pas

Une fourmi parlant français
parlant latin et javanais
ça n'existe pas, ça n'existe pas

eh ! et pourquoi pas !

(2) Une fourmi traînant un char, plein de pingouins et de canards, ça n'existe pas, ça n'existe pas. Une fourmi parlant français, parlant latin et javanais, ça n'existe pas, ça n'existe pas. Eh ! Et pourquoi pas !

Comme on peut le voir, la fin des strophes est systématiquement indiquée par un point même s'il n'y avait pas de ponctuation dans le texte original. La fin des vers est retranscrite par une virgule, sauf dans le cas où une ponctuation existait déjà.

3.2.2 Méthodologie

Les enregistrements audio ont d'abord été transcrits et segmentés en phrases. La transcription orthographique a ensuite été phonétisée, et le signal acoustique automatiquement segmenté en phones, syllabes, et mots. Les transcriptions phonétiques et les segmentations acoustiques ont été vérifiées et corrigées, si nécessaire. L'ensemble des données a été utilisé pour l'analyse rythmique et prosodique.

La voyelle plutôt que la syllabe a été choisie comme unité de base pour générer les patrons de durée et pour analyser et comparer les durées des pauses et les débits en fonction des genres et des locuteurs. Ce choix résulte du fait que les structures syllabiques varient beaucoup en français (entre, par exemple, des syllabes de forme CCVC et d'autres de forme CV). La durée des syllabes ne constitue donc pas un indicateur robuste pour évaluer les taux d'allongement. Comme le nombre de voyelles par contextes prosodiques était limité en raison de la taille du corpus, aucune normalisation des durées n'était possible. Aussi avons-nous décidé de faire une distinction entre voyelles courtes et voyelles longues, même si cette distinction n'existe pas dans le système phonologique du français. Les voyelles nasales ([ɔ̃],[ɑ̃], [ɛ̃] et [œ̃]) et les séquences composées d'une semi-voyelle et d'une voyelle en position de noyau (par exemple, [jɛ̃] dans *tiens* [tjɛ̃], [wa] dans *noir*

[nwaʁ]) ont été codées comme des voyelles longues, tandis que les autres voyelles ont été considérées comme courtes.

De nombreux travaux consacrés à la prosodie du français ont montré que l'intonation et l'accentuation sont très liées dans cette langue (cf. (Brechtje Post 2011)) ; aussi avons-nous décidé de partir des découpages prosodiques pour étudier les schémas rythmiques. Les différents textes ont donc été segmentés en groupes prosodiques, une distinction étant faite entre trois niveaux de structuration : le mot prosodique (MP) qui correspond à un mot lexical précédé des mots grammaticaux qui en dépendent, le syntagme phonologique (SP) qui est borné à droite par une tête lexicale de projection syntagmatique maximale et le groupe intonatif (IP). Pour pouvoir comparer les données malgré les différences possibles de réalisation et pour éviter une certaine circularité, nous avons décidé de dériver les unités prosodiques à partir du texte, et plus précisément des informations morpho-syntaxiques, voir (Elisabeth Delais-Roussarie 1996). De plus, comme la dernière syllabe des groupes prosodiques est considérée en français comme accentuée et est habituellement allongée, trois catégories de syllabes accentuées ont été retenues pour comparer les taux d'allongement par rapport à la position prosodique :

- AC-MP correspond à la dernière syllabe accentuée d'un mot prosodique ;
- AC-SP coïncide avec la dernière syllabe accentuée d'un syntagme phonologique ;
- AC-IP correspond à la dernière syllabe accentuée d'un IP.

3.3 Résultats

L'étude des durées observées pour les voix naturelles et synthétiques a permis de comparer les débits de parole, la durée et la distribution des pauses, et le marquage de la structure prosodique.

3.3.1 Débit de parole et pauses

La durée totale des textes lus a été utilisée pour calculer, pour chaque locuteur et pour chaque genre, le débit de parole, le taux d'articulation et les durées des pauses. Les différences entre débit de parole et taux d'articulation reposent sur le fait que les pauses ne sont pas prises en compte dans le second cas. Le tableau 3.5 présente les résultats obtenus pour chaque locuteur dans les trois genres. Pour chacun d'entre eux, les deux premières lignes concernent le débit de parole et le taux d'articulation, tandis que les deux dernières lignes portent sur la durée.

Les taux d'articulation et les débits de parole observés pour chaque genre varient de manière importante, mais on ne peut pas dire que les voix de synthèse diffèrent des voix naturelles : LOD et SY-P possèdent pour les trois genres considérés un débit plus rapide que les locuteurs SY-A, GOR et DRE (qui ont des débits plus lents, mais relativement semblables). Si on compare pour un genre donné les débits des différentes voix,

Comptines	LOD	DRE	GOR	SY-A	SY-P
Débit de parole moyen (ph./sec.)	9.9	7.35	7.08	7.63	9.09
Taux d'articulation moyen (ph./sec)	12.09	7.83	8.53	9.79	12.61
Durée totale des pauses (ms)	2178.92	1449.22	2573.76	3025	3000
% de pause moyen	25.27	13.60	24.15	29.06	33.80
Poèmes	LOD	DRE	GOR	SY-A	SY-P
Débit de parole moyen (ph./sec.)	10.6	8.16	6.28	8.26	9.45
Taux d'articulation moyen (ph./sec)	13.60	9.32	8.72	10.70	12.85
Durée totale des pauses	1534	1373.36	2590.52	2000	2000
% de pause moyen	27.38	18.10	33.85	28.17	31.29
Contes	LOD	DRE	GOR	SY-A	SY-P
Débit de parole moyen (ph./sec.)	10.58	8.74	8.18	9.31	10.79
Taux d'articulation moyen (ph./sec)	14.99	10.08	11.09	11.36	13.68
Durée totale des pauses	1331.33	763.40	1482.06	992	992.14
% de pause moyen	32.82	18.06	29.96	21.96	24.79

TABLE 3.5 – *Débit de parole et taux d'articulation en phones/sec, durée et pourcentage de pauses (relativement à la durée totale de lecture).*

on s'aperçoit que les taux d'articulation obtenus par la synthèse sont dans les limites de ceux observés pour les voix naturelles.

Une comparaison inter-genres montre que les locuteurs adaptent leur débit de parole et leur taux d'articulation en fonction du genre, des débits plus lents étant mis en œuvre pour la lecture des comptines et des poèmes. Cette adaptation est, comme on s'y attendait, moins claire pour la parole synthétique. De fait, pour une voix donnée et pour tous les genres, le même corpus et la même procédure de sélection d'unités sont utilisés. Néanmoins, les différences entre voix naturelles et voix synthétiques demeurent mineures, ce qui signifie que l'adaptation découle également de la composition interne des textes.

Concernant les pauses, une différence importante existe entre parole naturelle et parole synthétique dans tous les genres. La proportion de pauses est moins importante dans les comptines que dans les contes pour les trois locuteurs ; en revanche, il y a plus de pauses dans les comptines que dans les contes pour les deux voix de synthèse. Vus les mécanismes de placement des pauses utilisés par le synthétiseur, ces résultats sont tout à fait logiques. De plus, en parole naturelle, la durée des pauses semble dépendre du taux d'articulation (la proportion de pauses est en effet moins importante lorsque le débit est lent, comme par exemple dans les comptines et les poèmes) ; mais une telle corrélation n'apparaît pas en synthèse, une durée fixe étant assignée aux pauses en fonction de la force de la frontière prosodique.

Dans l'ensemble, on n'observe pas de grosses différences entre parole naturelle et synthétique pour le débit de parole et le taux d'articulation. En effet, les voix de synthèse et les

voix naturelles varient dans les mêmes proportions. En revanche, pour la durée et la proportion des pauses, il existe des différences entre la synthèse et les voix naturelles.

3.3.2 Structure prosodique et durée

En règle générale, les allongements indiquent en français le phrasé et l'accentuation. Les syllabes accentuées, qui correspondent à la dernière syllabe pleine à chaque niveau de structuration prosodique, sont allongées, leur taux d'allongement étant proportionnel à la force de la frontière prosodique. Aussi nous avons voulu vérifier si cela se retrouve dans les voix de synthèse. Pour ce faire, les taux d'allongement ont été calculés en comparant les durées des voyelles dans les syllabes non accentuées aux durées des segments vocaliques dans les syllabes accentuées, et cela à tous les niveaux de hiérarchie prosodique (mot prosodique, syntagme phonologique et groupe intonatif). Le tableau 3.6 présente les résultats obtenus par genre.

Il existe une variation relativement importante de la durée des voyelles non accentuées dans les différents genres, et cela pour les trois voix naturelles. De manière générale, les voyelles en position non accentuée sont plus longues dans les comptines et les poèmes que dans les contes. Par comparaison, aucune variation claire n'est observée entre genres pour les voix synthétisées. Ce résultat confirme le fait que les locuteurs adaptent leur débit de parole en fonction du genre, ce que ne fait pas la synthèse de parole.

En ce qui concerne le marquage de la structuration prosodique, des allongements se produisent toujours à la fin des groupes prosodiques à tous les niveaux (mot prosodique, syntagme phonologique et groupe intonatif), en synthèse comme en parole naturelle. Pour tous les genres et tous les locuteurs, les taux d'allongement varient :

- de 10 à 40%, au niveau du mot prosodique, avec une moyenne aux alentours de 20% ,
- de 20 à 100% au niveau du syntagme phonologique, avec une moyenne à 40%,
- de 60 à 190% au niveau du groupe intonatif, avec une moyenne de 98% (avec 77% en parole naturelle et 128% en synthèse).

Les taux moyens (à l'exception des IP en parole synthétique) correspondent à ceux souvent donnés dans les travaux sur les patrons de durée en français (Elisabeth Delais-Roussarie 1996). Dans les comptines, les taux d'allongement ne permettent pas toujours de clairement distinguer les trois niveaux de structuration, en particulier les mots prosodiques des syntagmes phonologiques. Dans les poèmes, la distinction entre SP et IP n'est pas clairement marquée dans les taux d'allongement chez LOD et GOR. On peut aussi noter que les taux d'allongement qui indiquent les frontières des IP sont plus nettement marqués en synthèse qu'en parole naturelle, dans tous les genres, et plus particulièrement chez SY-A.

Comptines	LOD	DRE	GOR	SY-A	SY-P
Durée moyenne voy. non accentuée	66 ms	130 ms	93 ms	81 ms	68 ms
Taux d'allongement AC-MP	30%	20%	20%	20%	30%
Taux d'allongement AC-SP	20%	20%	50%	30%	10%
Taux d'allongement AC-IP	90%	70%	70%	150%	60%
Poèmes	LOD	DRE	GOR	SY-A	SY-P
Durée moyenne voy. non accentuée	67 ms	110 ms	95 ms	78 ms	69 ms
Taux d'allongement AC-MP	10%	20%	40%	20%	20%
Taux d'allongement AC-SP	60%	40%	100%	50%	40%
Taux d'allongement AC-IP	60%	70%	80%	190%	80%
Contes	LOD	DRE	GOR	SY-A	SY-P
Durée moyenne voy. non accentuée	59 ms	99 ms	78 ms	77 ms	65 ms
Taux d'allongement AC-MP	10%	20%	10%	20%	20%
Taux d'allongement AC-SP	20%	40%	50%	40%	30%
Taux d'allongement AC-IP	80%	80%	100%	190%	100%

TABLE 3.6 – *Durées moyennes des voyelles dans les syllabes non accentuées (en ms.) et taux d'allongement (en %) pour les trois niveaux de structuration (MP, SP et IP).*

Globalement, les patrons de durée obtenus pour la parole synthétique sont relativement comparables à ceux observés en parole naturelle : les différents niveaux de phrasé sont toujours indiqués par un allongement dont le taux varie, très souvent, proportionnellement à la force de la frontière (Brechtje POST 2000 ; E DELAIS-ROUSSARIE et al. 2015).

3.4 Discussion

Les comparaisons effectuées ne révèlent pas de différences notables entre parole synthétique et parole naturelle. De fait, les variations qui apparaissent pour le débit de parole et le taux d'articulation ne permettent pas de distinguer la parole naturelle de la parole synthétique. En ce qui concerne les allongements et le marquage des frontières prosodiques, l'analyse montre clairement que des allongements sont réalisés en fin de groupement prosodique dans les deux types de parole (naturelle et synthétique), même si des différences apparaissent dans l'importance des taux d'allongement observés au niveau des IP (plus important en synthèse qu'en parole naturelle). Néanmoins, on peut douter que ces différences expliquent à elles seules le manque de naturel de la parole synthétique. De plus, en écoutant les stimuli synthétisés, nous avons été surpris par la qualité des patrons rythmiques observés dans les comptines, en particulier pour SY-A : ils semblent très naturels en comparaison de ceux obtenus pour les contes. En conséquence, les problèmes de rythme rencontrés en synthèse ne peuvent pas être attribués à des « sur-allongements » au niveau des IPs.

Les taux d'articulation et le débit d'une part, et le marquage par la durée de la structuration prosodique d'autre part, ne peuvent expliquer le manque de naturel dans les motifs rythmiques. Il faut donc trouver d'autres explications. Deux pistes de recherche méritent selon nous d'être explorées :

- aucune corrélation entre le débit de parole, la force des frontières et la durée des pauses n'a été observée en parole synthétique, alors que cette corrélation existe en parole naturelle (en français, de nombreuses études ont montré que les groupes prosodiques comme le syntagme phonologique ou le groupe intonatif visent soit à posséder le même nombre de syllabes soit la même durée (Elisabeth Delais-Roussarie 1996), les durées des pauses pouvant alors jouer un rôle important dans la recherche de l'isochronie) ;
- les patrons intonatifs jouent probablement un rôle dans la réalisation des motifs rythmiques : en insérant une virgule à la fin de chaque vers dans les poèmes et les comptines, on a forcé la réalisation d'un contour mélodique non final montant (contour de continuation majeure), ce qui a conduit à la répétition régulière d'une forme mélodique et a renforcé l'impression de rythme, ces résultats laissant penser que la récurrence de motifs mélodiques est cruciale pour le rythme.

3.5 Prédiction des groupes prosodiques pour la dictée

L'utilisation de procédures automatiques et de la synthèse de la parole dans le cadre d'une dictée impose plusieurs contraintes :

- l'intelligibilité de la parole synthétique doit être excellente afin que l'apprenant puisse entendre chaque mot en vue de son écriture ;
- les groupes prosodiques, tout en ayant une taille raisonnable, doivent permettre de gérer les difficultés orthographiques et grammaticales (règles d'accord, etc.) ;
- l'environnement doit tenir compte, notamment, de la vitesse de frappe de frappe de l'apprenant.

Dans cette section, nous nous focalisons tout d'abord sur la génération automatique de groupes à partir du texte de la dictée. Afin de proposer une procédure automatique, nous avons tout d'abord analysé un ensemble de dictées réalisées en primaire. Les résultats de ces observations ont servi à définir et formaliser les règles utilisées par l'algorithme de génération de dictée. Dans un second temps, l'évaluation de la méthode, réalisée auprès d'élèves et d'enseignants, est détaillée.

3.5.1 Méthode

Constitution du corpus

Afin d'étudier les découpages prosodiques et l'intonation dans le cadre de la lecture de dictée, un ensemble de dictées lues à des enfants allant du CP au CE2 en France et au Québec (Canada) a été collecté. Ces données proviennent de trois sources :

- quatre dictées courtes sont issues de l'association canadienne "Fondation Paul Gerin-Lajoie", en particulier, les dictées du premier niveau (équivalent CP) ;
- quatre dictées sont issues du site français *Ladictéee.fr* qui offre un ensemble varié de dictées et d'exercices grammaticaux pour les écoliers ;
- deux dictées ont été enregistrées en situation réelle par des chercheurs du projet Phorevox.

Le choix des dictées a été effectué afin de couvrir les différents niveaux scolaires visés pour le logiciel. Bien que le nombre de dictées étudiées soit limité, il est important de noter que les variations portent plus sur la façon de répéter les groupes et d'en introduire de nouveaux, que sur la formation des groupes à proprement parler. Ce point est essentiel, dans la mesure où c'est surtout la formation des groupes qui nous importe ici.

Les dictées ont été annotées par deux des auteurs en ajoutant, pour chaque texte, deux types d'informations : les découpages prosodiques (i.e. la manière dont les textes ont été segmentés en groupes durant la lecture) ; et la forme du contour intonatif observé à la fin des différents groupes prosodiques. L'annotation a été effectuée au moyen d'une analyse perceptive et d'une étude instrumentale des données. L'analyse perceptive a consisté à écouter attentivement le signal sonore, ce qui a permis de définir les groupes et la forme des contours intonatifs à la fin des groupes. L'analyse acoustique, réalisée en utilisant le logiciel Praat (Boersma 2002), a pour but de valider ce qui a été perçu. Une attention particulière a été portée aux contours intonatifs et aux pauses, qui indiquent en partie les découpages.

Méthodologie générale

Avant de décrire en détail les règles utilisées pour obtenir les groupes, ainsi que les procédures mises en place pour effectuer la répétition et introduire de nouveaux groupes, nous souhaitons faire mention de trois points qui ont été observés dans l'ensemble des dictées analysées. Tout d'abord, le titre et le texte sont lus intégralement une fois à un débit relativement lent avant la lecture de la dictée proprement dite. Ensuite, durant la phase de dictée, la phrase complète est répétée une fois dès que tous les groupes la composant ont été lus. Enfin, les marques de ponctuation sont prononcées aux positions auxquelles elles apparaissent dans le texte afin que l'élève puisse les écrire, comme cela est illustré sous (1) et (2), où la transcription orthographique de ce qui a été prononcé est indiquée.

(1) *avec Papa point je marche dans la nature avec Papa point*
(2) *à l'école virgule je travaille toujours avec lui point*

3.5.2 Découpages prosodiques observés

Dans les données observées, les groupes utilisés durant la dictée correspondent, dans plus de 95% des cas, à des syntagmes mineurs. Pour un groupe prépositionnel, par exemple, le nom est toujours dicté avec la préposition et le déterminant, comme indiqué dans (3), et de même, un auxiliaire est dicté avec le verbe comme dans (4).

(3) *Je marche dans la nature avec Papa* → (je marche)$_{MiP}$ (dans la nature)$_{MiP}$ (avec Papa)$_{MiP}$
(4) *Il se demande si sa maman a trouvé les bons médicaments* → (il se demande)$_{MiP}$ (si sa maman)$_{MiP}$ (a trouvé)$_{MiP}$ (les bons médicaments)$_{MiP}$

Il faut toutefois noter que deux syntagmes mineurs, obtenus grâce à l'information morphosyntaxique, peuvent être regroupés en un seul syntagme lorsque leur taille est inférieure à deux syllabes. Dans beaucoup de cas, par exemple, l'auxiliaire *être* est regroupé avec l'attribut comme cela est montré dans (5). Néanmoins, il arrive que cette règle de regroupement ne soit pas respectée, notamment si cela donne naissance à un groupe relativement long, comme sous (6).

(5) *Le lac est bleu* → [le lac] [est bleu]
(6) *C'est mon meilleur ami* → [c'est] [mon meilleur ami]

Chaque groupe prosodique mineur est réalisé comme un IP (c'est-à-dire suivi d'une pause), qu'il soit isolé comme sous (7) ou bien intégré à une phrase comme sous (8).

(7) *Avec mon ami.* → [avec mon ami]$_{IP}$ [point]$_{IP}$
(8) *Je joue dans l'eau avec mon ami* → [je joue]$_{IP}$ [dans l'eau]$_{IP}$ [avec mon ami]$_{IP}$ [point]$_{IP}$

Cela signifie qu'en dictée - vu ici comme un style de parole particulier - un groupe prosodique mineur est réalisé avec les caractéristiques prosodiques des IPs, en particulier avec un allongement final important et la présence d'une pause. Cette prosodie a été décrite comme complètement appropriée en français par Verluyten 1982. Il résulte de ce que Verluyten 1982 appelle *l'élasticité prosodique* c'est-à-dire le fait qu'un MiP peut être réalisé comme un IP sans restructuration supplémentaire.

La procédure de segmentation utilisée pour dicter le texte repose ainsi sur une analyse qui introduit une coupure majeure après les noms, verbes, adjectifs et adverbes s'ils ne sont pas des modifieurs du mot suivant. Ce dernier principe doit permettre de prononcer dans un même IP un adjectif prénominal et un nom, comme dans « le petit garçon », un adverbe et un adjectif comme dans « très ennuyeux », ou un auxiliaire et le participe passé comme dans « est arrivé ».

3.5.3 Procédures de répétition

En plus des règles de segmentation et de prononciation des groupes, il existe des procédures pour introduire les nouveaux groupes. Elles sont décrites dans les paragraphes suivants. Comme une certaine variabilité existe, nous avons inféré de l'observation des données trois procédures distinctes.

Procédure IP par IP

Cette procédure consiste à prononcer une phrase de la façon suivante. La phrase est d'abord prononcée en entier, mais chaque syntagme mineur MiP est séparé du suivant par une pause dont la durée varie entre 500 ms et 1 seconde. Puis chaque MiP est produit en isolation (voire parfois répété), une pause de 2 secondes minimum est alors réalisée entre un MiP et celui qui suit (ou la répétition de ce même MiP). Pour finir, une fois que tous les MiP ont été produits en isolation, la phrase entière est à nouveau répétée, chaque MiP étant alors séparé du suivant par une pause de 500 ms à 1 seconde. Généralement, les mouvements mélodiques réalisés à la fin des MiP sont montants. Pour (9), cette procédure amène au découpage et à la réalisation présentée en (10). Chaque retour à la ligne indique que le groupe est prononcé de manière séparée des autres et suivi d'une longue pause. Cette procédure a principalement été utilisée pour les dictées faites devant les plus jeunes élèves (1re année d'école élémentaire).

(9) Je joue dans l'eau avec mon ami.

(10) [je joue]$_{IP}$ [dans l'eau]$_{IP}$ [avec mon ami]$_{IP}$ [point]$_{IP}$
[je joue]$_{IP}$
[dans l'eau]$_{IP}$
[avec mon ami]$_{IP}$ [point]$_{IP}$
[je joue]$_{IP}$ [dans l'eau]$_{IP}$ [avec mon ami]$_{IP}$ [point]$_{IP}$

Procédure par chaînage des IP

La procédure par chaînage d'IP repose sur une segmentation en MiP, et sur l'introduction progressive des MiP, lors d'une répétition. Comme pour la procédure précédente, chaque IP correspond, pour ce qui est de son extension, à un MiP. Une phrase comme (11) est segmentée comme indiqué sous (12). Les sauts de ligne rendent compte de la présence de pauses très importantes, d'une durée supérieure à 3 secondes. L'IP [le chien], qui est comparable à un mot prosodique, est d'abord produit en isolation et suivi d'une longue pause. Il est alors répété, et suivi de l'IP (ou MiP) qui le suit. Une pause d'une durée courte (200 ms en moyenne) est réalisée entre les deux IP [le chien] et [s'étire]. L'IP introduit est alors réalisé en isolation et suivi d'une longue pause, et ainsi de suite. Une fois que la phrase a été produite, elle est répétée, la segmentation étant clairement marquée par la réalisation de pauses d'une durée de 200 à 500 ms.

(11) Le chien s'étire sur le tapis.

(12) [le chien]$_{IP}$
[le chien]$_{IP}$ [s'étire]$_{IP}$
[s'étire]$_{IP}$
[s'étire]$_{IP}$ [sur le tapis]$_{IP}$
[sur le tapis]$_{IP}$ [point.]$_{IP}$
[le chien]$_{IP}$ [s'étire]$_{IP}$ [sur le tapis]$_{IP}$ [point]$_{IP}$

Même si cette procédure a été utilisée dans approximativement 25% des cas dans nos données, elle possède de nombreuses faiblesses, en particulier, dans le cas de phrases complexes. Par exemple, lorsque le complément s'intercale entre le sujet et le verbe, l'accord ne peut être déduit de manière directe, ce qui peut induire des erreurs.

Procédure phrase par phrase

Cette dernière procédure consiste à dicter le texte phrase par phrase. Chaque phrase est prononcée une ou deux fois en fonction de sa taille, à un débit relativement lent, et elle est suivie d'une pause longue. Pour ce qui est de la segmentation interne, chaque MiP est séparé du MiP suivant par une pause dont la durée varie entre 200 et 600 ms. La présence de la pause permet de considérer les MiP comme équivalents à des IP.

Dans les cas complexes ou dans les cas de phrases longues, la segmentation peut être réalisée proposition par proposition. Une proposition réfère à plusieurs types d'éléments : (i) un syntagme séparé par une virgule, dans le cas d'un ajout en périphérie suivi d'une proposition comme la séquence soulignée sous (13) ; (ii) des propositions subordonnées ou coordonnées comme sous (14). Lorsqu'une telle procédure de découpage est utilisée, la phrase complète est prononcée une fois de plus après que toutes les parties ont été dictées.

(13) À l'école, je travaille toujours avec lui.
(14) Mon père rentre très tard à la maison parce qu'il est musicien.

3.5.4 Algorithme de découpage et répétitions

L'algorithme se décompose en trois étapes. Tout d'abord, une analyse syntaxique du texte est réalisée. Les mots du texte correspondant à des feuilles de l'arbre syntaxique sont regroupés selon leur proximité dans l'arbre, et en tenant compte des contraintes de taille minimale (en terme de nombre de mots par groupe). En raison des erreurs possibles d'analyse et des contraintes de taille, le système permet de générer des groupes réduits à un seul mot comme « comme » sous (15) et des groupes possédant une ponctuation au milieu comme « En réalité, c'est » sous (16).

(15) *Avec une cheminée qui crache de la fumée, comme une locomotive à vapeur.*
→ (avec une cheminée) (qui crache) (de la fumée) (comme) (une locomotive) (à vapeur)

(16) *En réalité, c'est un hélicoptère* → (en réalité, c'est) (un hélicoptère)

Afin d'éviter cela, les marques de ponctuation sont considérées comme des frontières prosodiques. Ainsi, l'énoncé (16) est divisé comme suit : *(en réalité,) (c'est un hélicoptère).*

De plus, les groupes ne contenant qu'un seul mot sont fusionnés avec un groupe adjacent. Ainsi, « comme » est regroupé avec « une locomotive ». Le mécanisme de regroupement repose sur le nombre de syllabes par groupe comme paramètre principal, et non sur le seul nombre de mots. De plus, le nombre de voyelles graphiques non consécutives est pris en compte comme estimateur du nombre de syllabes. Pour le test, chaque groupe contient au moins 5 voyelles non consécutives. À l'issue du découpage, une indication de contour mélodique est ajoutée : un contour montant à la fin d'un groupe et avant une ponctuation ; un contour descendant avec un point. Enfin, des pauses courtes sont ajoutées entre un mot et une ponctuation tandis que des pauses longues (environ 3s) sont insérées après chaque groupe.

Une fois le texte segmenté en groupes, il est possible de le dicter de manière automatique. Chaque groupe est ensuite annoté de sorte qu'il soit prononcé de manière isolée (suivi par une pause très longue), les marques de ponctuation sont rendues explicites. Le texte annoté est ensuite fourni en entrée du système de synthèse qui peut alors le traiter, en utilisant une voix de synthèse conçue de manière spécifique pour la dictée.

3.5.5 Méthodologie d'évaluation

L'évaluation est réalisée auprès d'enseignants et d'élèves et consiste en une évaluation subjective de dictées produites de manière automatique. Cette évaluation porte sur trois points : la segmentation du texte en groupes de mots en vue de la dictée (emplacement des frontières et taille des groupes) ; le débit de la voix synthétisée utilisée pour les dictées ; et les patrons prosodiques associés aux groupes dictés. Dans ce paragraphe, nous présentons les stimuli à partir desquels les questionnaires sont construits (un pour chaque population), les participants, ainsi que la procédure suivie.

Stimuli utilisés

Les stimuli présentés sous (17) et (18) sont deux textes extraits d'un livre d'exercices pour enfants. Comme mentionné auparavant, plusieurs découpages peuvent être générés par l'algorithme en fonction des paramètrages retenus. Pour le test, qui porte sur la tâche de dictée, nous avons retenu les découpages donnés sous (17) et (18) et indiqués par le symbole "—".

(17) *Lila marche sur la plage. — Elle voit un crabe qui — se cache sous une algue. — Elle soulève l'algue avec — un bout de bois. — Mais le crabe n'est plus — là, — il a disparu dans le sable. —*

(18) *Quand Marinette joue — de la trompette, — Marion joue du violon, — Martin tape sur un tambourin, — Mario gratte son banjo, — Manon nous chante une chanson, — et moi, — je frappe dans les mains. —*

Ces stimuli ont été choisis pour deux raisons principales. La première est que le découpage généré présente des erreurs qui apparaissent fréquemment et qui peuvent s'avérer problématiques pour la tâche de dictée (certaines frontières, par exemple, ne vont pas dans le sens des découpages grammaticaux, comme « l'algue avec »). La deuxième raison est que nous souhaitions évaluer des groupes, créés avec des paramètres de taille que nous jugions à priori trop longs, notamment pour le public enfant.

Pour chaque dictée, le texte complet est lu une première fois avec un débit normal, la dictée commence ensuite. La voix de synthèse spécifique à la dictée produit des pauses entre chaque groupe. La durée des pauses peut être fixée par l'utilisateur.

Participants

L'outil que nous développons est dédié à un public d'enfants d'écoles primaires en cours d'acquisition de l'écriture et de la lecture (cycle 2, CP et CE1), et il doit offrir des exercices en lien avec les difficultés des apprenants à l'écrit. Il était donc important de savoir ce que les utilisateurs potentiels, les élèves et les enseignants d'école primaire, pensent de la procédure de dictée.

Un groupe de 7 filles ayant entre 6 ans 1 mois et 8 ans 2 mois et vivant en région parisienne a participé au test. Parmi ces élèves, une était en CE2 (8 ans 2 mois lors du test), une autre était en CP (6 ans 1 mois), et 5 en CE 1 (l'âge moyen étant de 7 ans 2 mois). Les élèves ont été choisies par les enseignants dans la mesure où elles n'avaient aucune difficulté scolaire, et présentaient même une certaine aisance dans l'acquisition de la langue écrite.

Le groupe des enseignants est quant à lui composé de 14 personnes (6 vivant en Région Parisienne, 6 en Bretagne, et 2 en Pays de Loire) qui ont accepté de participer au test de manière gratuite. Ils sont tous en activité comme enseignant dans des écoles primaires, à l'exception d'un participant qui est orthophoniste. Parmi eux, 10 sur 14 enseignent en cycle 2 (CP-CE1), et certains sont très intéressés par l'usage de nouvelles technologies dans leur classe même si seuls deux participants ont déjà utilisé des outils pédagogiques incluant de la synthèse de parole.

Procédure d'évaluation

La procédure d'évaluation est adaptée à chaque groupe. En conséquence, deux questionnaires ont été réalisés afin de collecter des informations sur le découpage et sur la tâche elle-même.

L'évaluation par les enfants est effectuée de la manière suivante. Les enfants doivent réaliser la dictée. Pour des raisons de place, nous ne présenterons ici que les résultats pour la première dictée (voir (17)). Nous avons utilisé un ordinateur portable équipé de haut-parleurs afin de lire la dictée produite par le système de synthèse. Les groupes ont été produits comme indiqué dans le paragraphe 3.5.4. Aucune limite de temps n'a été imposée pour la réalisation de la dictée, et chaque groupe pouvait être répété autant de fois que nécessaire. Pendant la dictée, l'interviewer a pris note des emplacements pour lesquels la compréhension était difficile ainsi que du nombre de répétitions de chaque groupe. Après la dictée, les enfants devaient répondre oralement au questionnaire (en majorité des questions de type oui/non et quelques questions libres). Ces questions concernaient (i) l'intelligibilité de la parole synthétique, en particulier lorsque des groupes ont été oubliés par les enfants ; (ii) les tailles des groupes, en particulier dans le cas des répétitions ; (iii) l'usage d'une voix de synthèse pour la dictée.

L'évaluation par le groupe des enseignants consiste en un test en ligne dans lequel les participants écoutent les stimuli (sans aucune contrainte), répondent aux questions, et peuvent donner leur opinion sur différents aspects de l'outil. Le questionnaire est divisé en trois parties : la première concerne le stimulus (17), la seconde concerne le stimulus (18) et la troisième présente des questions sur le débit de parole, et sur l'adéquation des patrons mélodiques dans une séquence particulière. Les participants de ce groupe écoutent tout d'abord la dictée, puis lisent le texte de la dictée présenté à l'écran. Ils répondent alors à des questions qui adressent les points suivants : la localisation des frontières de groupes, en particulier lorsqu'une erreur de découpage apparaît, la taille des groupes et l'intelligibilité de certains groupes. Afin d'être en mesure d'interpréter les résultats, la plupart des items du questionnaire attendent des réponses fermées (de type oui/non, trop lent/ normal/ trop rapide, etc.). Les résultats bruts de l'évaluation (et non en pourcentage en raison du nombre réduit de participants) sont présentés et discutés dans la section suivante.

3.5.6 Résultats

Évaluation par les enfants

Le tableau 3.7 donne les résultats pour chaque item du questionnaire concernant la voix de synthèse pour la dictée. On peut observer en premier lieu que les enfants peuvent réellement être des utilisateurs potentiels de l'application puisqu'ils trouvent cet outil utile dans le cas des dictées. La voix de synthèse est jugée naturelle malgré le fait que trois élèves trouvent la manière de prononcer étrange. La plupart des enfants ont éprouvé des difficultés à comprendre les mots « algue » (pour la deuxième occurrence lorsqu'il est groupé avec « avec »), « crabe » (à chaque fois) ainsi que le prénom « Lila ». De plus, 6 enfants sur 7 trouvent le débit trop élevé pour la dictée même si aucun des élèves n'a eu de difficulté à comprendre les phrases dictées.

Concernant les frontières, les élèves ne semblent pas gênés par le fait que « crabe » soit

	Oui	Non
(a) J'aime les dictées	6	1
(b) Je voudrais utiliser un outil comme celui-ci pour faire des dictées à la maison ou en classe	7	0
(c) La voix utilisée dans la dictée ressemble à celle d'un robot	1	6
(d) La voix est agréable	7	0
(e) La manière de lire la dictée est étrange	4	3
(f) Le sens de la phrase est compris [1]	7	0
(g) La dictée était facile	3	4
(h) La manière de lire était trop rapide	6	1
(i) Tous les mots étaient faciles à comprendre	2	5
(j) C'est bizarre de regrouper « crabe qui »	2	5
(k) Le groupe « Lila marche sur la plage » est trop long	7	0
(l) C'est bizarre de regrouper « l'algue avec »	7	0
(m) Le groupe « il a disparu dans le sable » est trop long	6	1
(n) Mots difficiles à comprendre : algue (6), crabe (4), Lila (4), bout de bois (1)		

TABLE 3.7 – *Résultats du questionnaire enfants. Scores bruts (de (a) à (m)) et réponses libres pour la question (n). Le nombre d'enfants ayant mentionné le mot est donné entre parenthèses.*

regroupé avec le pronom « qui ». Seulement deux enfants ont trouvé l'absence de frontière à cet endroit gênant. Cependant, regrouper « l'algue » avec la préposition « avec » a été perçu comme étrange, mais cela est probablement dû à l'intelligibilité du mot « algue » à cet endroit.

Enfin, les élèves trouvent que les groupes constitués d'un énoncé complet comme « Lila marche sur la plage » ou « il a disparu dans le sable » sont trop longs. Chaque groupe a été répété $7,7$ fois en moyenne. En considérant les différents groupes, nous observons que la longueur des groupes n'est pas liée au nombre de répétitions (sauf pour le groupe 7 qui se résume au monosyllabique « là »). Nous notons également que le nombre de répétitions ne reflète pas les difficultés des enfants puisque la plupart ont demandé des répétitions dans le but de vérifier ce qu'ils avaient écrit. Les erreurs commises par les élèves sont cohérentes avec les mots perçus difficiles à comprendre. Ainsi, les groupes incluant les mots « crabe », « algue » et « Lila » sont les groupes qui possèdent le plus d'erreurs (pour 4 enfants sur 7).

Pour résumer, les enfants considèrent les groupes assez longs mais ne semblent pas gênés par l'absence de correspondance entre frontières syntaxiques et prosodiques. Ils adoptent plutôt une stratégie de dictée mot par mot. Par exemple, dans le cas de « là », tous les enfants ont dans un premier temps écrit « la » sans placer d'accent, puis quatre d'entre eux ont corrigé le mot.

	Oui	Non
(a) Trouvez-vous gênant que le pronom relatif « qui » soit regroupé avec son nom référent et non sa subordonnée ?	10	4
(b) Pensez-vous que ce groupe peut induire les enfants en erreur ?	10	4
(c) Pensez-vous que de tels groupes sont un obstacle à l'usage du logiciel ?	10	4
(d) Trouvez-vous gênant que la préposition « avec » soit détachée de son complément ?	10	4
(e) Si vous avez répondu Oui en (4), pensez-vous que la préposition doit toujours être regroupée avec son complément ?	10	0
(f) Trouvez-vous trop long le groupe « Lila marche sur la plage » ?	8	6
(g) Trouvez-vous trop long le groupe « il a disparu dans le sable point » ?	9	5
(h) La séquence « mais le crabe n'est plus là » est scindée en deux groupes. Auriez-vous fait la même chose pendant une dictée ?	5	9
(i) Pensez-vous que le groupe « là virgule » est trop court ?	10	4

TABLE 3.8 – *Résultats pour le questionnaire enseignants, première dictée (17).*

Évaluation par le groupe d'enseignants

Le tableau 3.8 [2] donne les résultats du questionnaire concernant la première partie de l'évaluation. Contrairement aux enfants, le groupe des adultes semble plus attaché au respect des frontières liées aux constructions syntaxiques. 10 participants parmi 14 considèrent qu'un pronom relatif doit être regroupé avec la subordonnée qu'il introduit, et qu'une préposition comme « avec » doit être regroupée avec le groupe nominal qu'elle introduit. Les découpages proposés par les enseignants montrent que les déterminants doivent être regroupés avec leur nom (à l'exception de 2 réponses parmi 13, pour lesquelles une frontière est proposée entre « un » et « tambourin »). De manière générale, les enseignants considèrent que le découpage doit respecter les frontières morphosyntaxiques ; dans le cas contraire, cela peut induire les élèves en erreur. L'évaluation par les enfants a montré que c'était bien le cas pour la préposition « avec » mais pas pour le pronom relatif. Des études plus systématiques sur ce point sont nécessaires afin de voir si la position des enseignants se fonde sur des résultats effectifs d'élèves, ou si elle repose sur une conception normative de la langue. Dans le cas de la séquence « mais le crabe n'est plus là », la plupart des participants ont proposé une frontière juste après le groupe sujet (11 cas sur 14, et pour un cas seulement, une frontière après l'adverbe « mais »). Cependant, l'évaluation de la seconde dictée (cf. tableau 3.9) montre que les enseignants acceptent de regrouper le sujet uniquement avec le verbe (9 cas sur 14) et non avec le complément du verbe, même si du point de vue de la syntaxe et de la prosodie en français, on fait souvent l'hypothèse d'une frontière forte entre le sujet et le verbe . Seulement 50% des participants auraient préféré une frontière après le sujet, et 8 participants sur 14 ne pensent pas que le sujet doive toujours être regroupé avec le verbe.

2. Par manque de place, les réponses libres sur les découpages des deux séquences sont données dans le texte.

L'examen des groupes proposés par les enseignants confirme cette remarque lorsque le sujet est un nom : 6 participants sur 8 ont formé le groupe (Lila marche) [3] et 10 sur 13 le groupe (Marion tape). Ce regroupement est systématique lorsque le sujet est un pronom (comme dans « il a disparu »). La séquence « elle soulève l'algue » est toujours préférée en un seul groupe, mais les enseignants peuvent être influencés par ce qu'ils ont entendu auparavant lorsqu'ils recommandent fortement de séparer le complément direct du complément indirect introduit par la préposition « avec ». Enfin, les découpages proposés par les enseignants montrent que les séquences courtes comme « mais », « et moi » et « là » peuvent former des groupes indépendants.

Concernant la taille des groupes, il semble que différentes stratégies puissent être adoptées. Ainsi, 10 participants sur 14 trouvent la séquence « là virgule » trop courte, mais « et moi » est accepté de manière unanime. Dans les autres cas, les réponses sont moins homogènes. Pour « Martin tape sur un tambourin », la séquence est perçue comme trop longue pour former un seul groupe (13 sur 14) mais les groupes « Lila marche sur la plage » et « il a disparu dans le sable » (ils ont respectivement 2 syllabes et 1 syllabe de plus que le premier groupe) sont acceptés comme pouvant former un seul groupe (resp. 6 et 5 réponses). Il est difficile de dire si c'est le nombre de syllabes, ou bien si c'est le tri-syllabique « tambourin » qui sont en cause. Ce dernier peut être considéré comme difficile et ainsi devrait être présenté dans un groupe plus court.

La dernière partie du questionnaire concerne le débit de parole et les patrons prosodiques utilisés dans des séquences considérées comme problématiques. Dans le cas du débit de parole, pour toutes les questions, une nette préférence apparaît pour l'emploi d'un débit adapté (13 sur 14) plutôt qu'un débit normal jugé trop rapide par les participants (12 sur 14). Rappelons que les enfants ont également trouvé le débit trop élevé. Le contour prosodique (montant ou descendant) ne semble pas avoir d'influence positive sur la séquence. Pour la séquence « Quand Marinette joue », 12 participants n'ont pas de préférence quant au patron prosodique utilisé. Pour des séquences problématiques, comme « elle soulève l'algue avec », nous obtenons des réponses assez aléatoires avec 5 pour un contour montant, 3 pour un contour plat et 6 pour un contour descendant. Il apparaît donc que le choix du contour intonatif utilisé n'améliore pas la voix de synthèse dans cette séquence particulière. Il semblerait même que certains cas ne suivent aucune règle en vigueur en français standard.

En résumé, comparées à celles des élèves, les réponses des enseignants soulignent l'importance de la syntaxe dans la détermination des frontières de groupe, au moins au niveau syntaxique le plus fin (i.e. les prépositions doivent être regroupées avec leur nom, les pronoms relatifs avec leur proposition, les déterminants avec leur nom, etc.). Cependant, pour les frontières après le groupe sujet, les réponses semblent être plus aléatoires et refléter des stratégies personnelles mises en oeuvre pendant la dictée. C'est également le cas pour la taille des groupes, où non seulement le nombre de syllabes est pris en compte mais aussi d'autres facteurs tels que les difficultés introduites par le lexique. Le débit

3. Seulement 8 participants ont donné une réponse à cette question.

	Oui	Non
(a) Trouvez-vous étrange de regrouper sujet et verbe dans « Marinette joue »	5	9
(b) Auriez-vous préféré deux groupes avec une frontière après le sujet comme dans (Quand Marinette) et (joue de la trompette) ?	7	7
(c) Pensez-vous que le sujet doit toujours est regroupé avec le verbe ?	6	8
(d) Si vous avez répondu Oui à la question (3), pensez-vous que des propositions subordonnées ou indépendantes doivent être prononcées en un seul groupe, même s'il est long ?	1	5
(e) Trouvez-vous le groupe « Martin tape sur un tambourin » trop long ?	13	1
(f) Trouvez-vous le groupe « et moi » trop court ?	0	14
(g) Dans une situation réelle, auriez-vous prononcé « Et moi, je frappe dans les mains » en un seul groupe ?	1	13

TABLE 3.9 – *Résultats pour le questionnaire enseignants, deuxième dictée (18).*

de parole utilisé pendant la dictée est généralement jugé satisfaisant par les enseignants et le type de contour prosodique ne semble pas améliorer l'intelligibilité des séquences problématiques.

3.5.7 Discussion

Dans ce travail, nous avons présenté un algorithme de découpage en groupes prosodiques permettant d'établir des groupes proches de ceux effectués lors d'une dictée. Une évaluation de la pertinence d'un algorithme de découpage en groupes prosodiques a également été présentée à partir de questionnaires adaptés aux élèves et aux enseignants. Ces derniers nous ont permis de mettre en avant les paramètres qui doivent être pris en compte pour améliorer, en particulier, le placement des frontières et la taille des groupes. Les résultats montrent des différences de jugement entre les élèves et les enseignants. Ils montrent également qu'un contrôle de la procédure par l'utilisateur semble nécessaire.

3.6 Conclusion

Dans ce chapitre, les travaux effectués en lien avec l'adaptation de modèles prosodiques pour la synthèse de parole ont été détaillés. Ainsi, une étude sur la différence entre quatre styles de parole (dictée, discours politique, roman et contes) a été présentée. Plusieurs facteurs ont été analysés ce qui a permis d'établir un ensemble de règles possibles pour obtenir des caractéristiques prosodiques pertinentes dans les styles de parole étudiés.

Ensuite, des travaux sur le rythme ont été menés dans le but de comprendre quels sont les défauts de la synthèse de parole et les différences avec la parole naturelle. Les résultats montrent que pour les trois genres littéraires étudiés, les différences perçues ne peuvent pas être expliquées seulement par la durée. Une hypothèse est alors que la longueur et la distribution de pauses, ainsi que les contours intonatifs choisis ont un impact important sur la perception du rythme et expliquent les différences relevées. Des travaux supplémentaires doivent être menés afin de vérifier cette hypothèse et proposer des moyens d'améliorer le cohérence de ces différents éléments.

Enfin, une application de l'adaptation de contraintes prosodiques pour la dictée a été présentée. Dans ce cadre, un algorithme de découpage en groupes prosodique a été implanté et évaluer dans des classes de cycle 2 auprès d'enseignants et d'élèves. L'algorithme a été à cette occasion intégré dans un plateforme d'apprentissage du Français réalisée dans le cadre du projet ANR Phorevox. Les résultats de l'évaluation montrent que l'adaptation de la synthèse de parole à un style particulier, comme la dictée, est possible, même s'il reste des imperfections pour ce qui est de l'algorithme de découpage en groupes prosodiques.

Afin de poursuivre ces travaux, des facteurs supplémentaires doivent être étudiés, comme ceux liés à la qualité vocale, qui possèdent une contribution importante à l'expressivité. De plus, des études sur le rythme sont en cours à travers la collaboration avec Elisabeth Delais-Roussarie dans le cadre du projet ANR SynPaFlex. Dans ce même cadre, nous sommes en train de construire un corpus de parole mono-locuteur de grande taille (environ 80h) contenant des livres audio. Ce corpus permettra de mener des études sur le rythme, les pauses ainsi que les contours intonatifs utilisés. Enfin, l'usage d'un tel corpus mono-locuteur permet également d'étudier les différences selon plusieurs axes : personnages, émotions, narration/dialogue.

Chapitre 4

Adaptation du moteur de synthèse de parole

Les travaux présentés dans ce chapitre ont été conduits principalement dans le cadre de la thèse de David Guennec, que j'ai dirigée. Les publications suivantes en sont le résultat : Guennec *et* Lolive *2014a ;* Guennec *et* Lolive *2014b ;* Guennec, Chevelu *et* Lolive *2015 ;* Guennec *et* Lolive *2016a ;* Guennec *et* Lolive *2016b. De plus, grâce à ces travaux, nous avons pu participer ces deux dernières années au challenge Blizzard (*Alain *et al. 2015 ;* Alain *et al. 2016).*

Trois principales familles de techniques existent en synthèse de parole. La première est l'approche paramétrique qui repose sur l'utilisation d'un modèle permettant la génération de paramètres représentant la parole ou plus récemment directement des échantillons de parole. Cette première famille regroupe les systèmes de synthèse paramétriques statistiques comme HTS (J. Yamagishi, Ling et Simon King 2008) ou les approches reposant sur des réseaux de neurones. La deuxième famille est celle des approches par concaténation de segments de parole, largement utilisée et implantée à l'heure actuelle, à titre d'exemple dans Alan W Black et Taylor 1994 ; A. J. Hunt et Alan W. Black 1996 ; Taylor, Alan W Black et Caley 1998 ; Breen et Jackson 1998 ; Clark, Richmond et Simon King 2007. La troisième combine les deux approches précédentes afin d'en tirer le meilleur et former des systèmes dits hybrides. Même si la majorité des travaux de recherche actuels se focalise sur les approches statistiques, ce sont les approches par concaténation et hybrides qui sont les plus largement utilisées dans les systèmes commerciaux. Dans tous les cas, le principal problème de recherche est le contrôle de l'expressivité lors de la synthèse (Rebordao et al. 2009).

Dans ce chapitre, le système de synthèse de parole par concaténation de l'équipe Expression, ses évolutions, ainsi que différentes propositions permettant d'améliorer la qualité de la synthèse de parole sont présentés. La première d'entre-elles concerne l'ajout de contraintes phonologiques lors de la sélection des unités afin d'éliminer les artéfacts liés

à la concaténation en favorisant des segments de parole supposés plus robustes. La seconde proposition concerne la possibilité d'introduire des contraintes prosodiques afin d'améliorer le contrôle de la prosodie. Enfin, les évaluations du système, avec différentes langues, lors des participations au challenge *Blizzard* seront présentées.

4.1 Architecture du système de synthèse de l'IRISA

Le principe de la sélection d'unités est de considérer un répertoire de segments de parole à partir desquels, un algorithme va chercher la séquence de segments optimale, au sens d'une fonction de coût, afin de créer un signal de parole correspondant à une séquence de phonèmes cible. La forme générale de la fonction de coût utilisée est classiquement présentée de la manière suivante (Alan W Black et Taylor 1994) :

$$U^* = \underset{U}{argmin} \left(\sum_{n=1}^{card(U)} C_t(u_n) + \sum_{n=2}^{card(U)} C_c(u_{n-1}, u_n) \right) \tag{4.1}$$

où U^* est la meilleure séquence de segments de parole, chaque u_n est une unité candidate dans la séquence U. $C_t(u_n)$ représente le coût de sélection du segment u_n par rapport à la cible recherchée. $C_c(u_{n-1}, u_n)$ permet de prendre en compte le coût de la concaténation de deux segments candidats et ainsi le risque d'obtenir des artéfacts.

Dans cette section, nous présentons tout d'abord l'architecture du système, puis nous nous intéressons aux différents blocs du moteur de synthèse de l'IRISA, développé dans l'équipe Expression. Ces différents éléments sont les filtres de pré-selection et les fonctions de coûts de sélection et de concaténation.

4.1.1 Architecture générale

Toutes les briques du système de synthèse sont interfacées avec la librairie Roots (Boëffard, Laure et al. 2012 ; Chevelu, Lecorvé et Lolive 2014a) qui permet de gérer de manière cohérente toutes les informations nécessaires au processus de synthèse. Cela permet également d'obtenir un système modulaire qui s'appuie sur des outils à l'état de l'art, par exemple pour la phonétisation (Duddington 2012) ou le POS tagging (Toutanova et Manning 2000). Les étapes de normalisation du texte et de syllabation sont quant à elles effectuées par des outils développés dans l'équipe.

Une fois l'analyse effectuée, la requête, constituée de la séquence de phonèmes cible et des consignes prosodiques, est transmise à la partie arrière du moteur de synthèse. C'est là qu'est réalisée la synthèse proprement dite, par l'utilisation de la sélection d'unités puis par une étape de génération du signal.

Dans la plupart des cas, la recherche de la meilleure séquence d'unités est réalisée grâce à un algorithme de Viterbi 1967. C'est le cas par exemple dans (A. J. Hunt et Alan W.

Black 1996 ; Conkie et al. 2000 ; Clark, Richmond et Simon King 2007). D'autres approches existent, comme l'utilisation d'un algorithme génétique dans (Kumar 2004). En pratique pour améliorer l'efficacité de la recherche, l'algorithme de Viterbi est implanté comme un algorithme de recherche en faisceau - *Beam Search* - en plaçant une contrainte sur le nombre de candidats à explorer pour chaque segment cible.

La recherche de la meilleure séquence d'unités peut être formulée comme un problème de recherche de meilleur chemin dans un graphe, nous proposons l'utilisation d'un algorithme de type A* (Guennec et Lolive 2014a). Cet algorithme permet sous certaines conditions de garantir l'optimalité de la solution tout en introduisant la possibilité d'améliorer la recherche par l'usage d'heuristiques.

Dans notre cas, la recherche de la meilleure séquence peut être effectuée grâce à l'algorithme A* ou Viterbi. Enfin, le signal de parole est généré avec une variante de TD-PSOLA (Moulines et Charpentier 1990).

4.1.2 Filtres de pré-selection

Une pré-sélection des unités est mise en place pour accélérer le processus de recherche d'unités de taille variable dans le corpus de parole (Conkie et al. 2000). Elle repose sur l'usage de filtres binaires portant sur des informations de nature linguistique, phonétique et prosodique.

L'implémentation que nous avons choisie ici repose sur une liste de filtres ordonnés par priorité. Ainsi, si le nombre de candidats u_n pour une unité cible t_n est trop faible, inférieur à MIN_u, le filtre le moins prioritaire est relâché. Autant de filtres que nécessaire sont relâchés afin d'obtenir un nombre suffisant de candidats. Le jeu de filtres utilisé varie notamment en fonction du choix corpus de parole et de la langue. Par exemple, pour le français, les filtres suivants peuvent être utilisés, du plus prioritaire au moins prioritaire :

1. Label du segment associé, diphonème ou autre (ne peut être relâché).
2. Est-ce un *Non Speech Sound* (ne peut être relâché) ?
3. Le phone est-il dans la dernière syllabe de la phrase ?
4. Le phone est-il dans la dernière syllabe du groupe de souffle ?
5. La syllabe courante est-elle en fin de mot ?
6. La syllabe courante porte-t-elle un contour intonatif montant ?
7. La syllabe courante porte-t-elle un contour intonatif descendant ?

Plus formellement, on considère que l'on dispose d'un n-uplet de J filtres modélisés par des fonctions indicatrices $f_j(u_n, t_n)$ $(j \in [0\,;J])$ valant 1 si u_n respecte la condition posée par le filtre j pour le diphone cible t_n et 0 sinon. Considérons l'ensemble des unités

satisfaisant les I premiers filtres pour le diphone cible t_n :

$$O(I_n, t_n) = \left\{ u_n / \prod_{i=1}^{I_n \leq J} f_i(u_n, t_n) = 1 \right\}. \tag{4.2}$$

L'étape de présélection vise à rechercher, pour chaque diphone cible t_n, l'ensemble $O(I_n, t_n)$ des noeuds candidats, à intégrer dans le graphe de sélection, pour lequel I_n est maximal et :

$$card(O(I_n, t_n)) \geq \text{MIN}_u. \tag{4.3}$$

Il est important de noter que l'ordre des filtres utilisés peut avoir un impact important lors de la sélection. La principale raison de ne pas intégrer ces filtres à la fonction de coût elle-même est de réduire la taille du graphe d'unités candidates (donc de réduire le temps de sélection). Cependant, il ne faut pas perdre de vue le fait qu'ils font partie intégrante du coût de sélection. En effet, les filtres constituent un ensemble de fonctions binaires de coût cible se fondant sur l'hypothèse suivante : si une unité ne respecte pas l'ensemble des filtres actifs, elle ne peut pas être utilisée pour la sélection.

On pourrait faire valoir que plus de filtres permettrait une meilleure sélection, mais par expérience plus de raffinement dans les filtres ne s'avère pas donner de meilleurs résultats. En effet, la meilleure unité est essentiellement un compromis entre un bon coût cible et un bon coût de concaténation, ce qui implique que chaque coût doit disposer d'un panel de choix suffisants.

4.1.3 Coûts de sélection et de concaténation

Le coût d'un segment candidat est généralement exprimé comme dans l'équation (4.1) et se divise en deux parties : le coût de sélection et le coût de concaténation. Dans ce paragraphe, nous introduisons ces deux sous-coûts et la forme qu'ils prennent dans le moteur de synthèse.

Coût de sélection Le coût de sélection évalue la proximité d'un segment candidat avec l'unité souhaitée dans la séquence cible. Il prend en compte des caractéristiques liées à l'identité du segment ainsi qu'à son contexte en termes de phonèmes voisins, de position dans la syllabe, dans le mot, dans le syntagme ou encore dans la phrase. Le coût de sélection, ou coût cible, prend ici la forme d'une somme pondérée de sous-coûts binaires, à la manière de Andrew J. Hunt et Alan W. Black 1996. Des exemples de critères utilisés dans le coût de sélection sont présentés dans le tableau 4.1.

Coût de concaténation Le coût de concaténation a pour but d'évaluer la proximité acoustique de deux segments candidats pour évaluer le risque de créer un artéfact audible dans le cas de leur concaténation. De nombreuses manières existent afin d'évaluer cela.

Text related features :	
TEXT_DIALOG	
Phoneme position :	
LAST_OF_BREATHGROUP	
LAST_OF_WORD	LAST_OF_SENTENCE
IN_CODA	IN_ONSET
SYLLABLE_BEGIN	SYLLABLE_END
WORD_BEGIN	WORD_END
Phonological features :	
LONG	NASAL
LOW_STRESS	HIGH_STRESS
Syllable related features :	
HAS_CODA	
LAST_SYL_OF_SENTENCE	
LAST_SYL_OF_BREATHGROUP	
SYLLABLE_RISING	SYLLABLE_FALLING

TABLE 4.1 – *Exemple de critères utilisés dans le coût de sélection pour l'anglais dans le cadre du challenge Blizzard.*

Nous considérons ici uniquement trois sous-coûts portant sur les coefficients cepstraux, ici MFCC, l'amplitude du signal et la valeur de la fréquence fondamentale F_0. Les valeurs utilisées pour calculer ces trois sous-coûts sont normalisées de manière à leur donner une importance égale. Ainsi, nous utilisons le coût de concaténation suivant :

$$C_c(u,v) = W_{\text{mfcc}}(u,v)C_{\text{mfcc}}(u,v) + W_{\text{amp}}(u,v)C_{\text{amp}}(u,v) + W_{\text{F0}}(u,v)C_{\text{F0}}(u,v) \quad (4.4)$$

où u et v sont deux segments candidats, $W_{\text{mfcc}}(u,v)$, $W_{\text{amp}}(u,v)$ et $W_{\text{F0}}(u,v)$ sont les coefficients de normalisation, respectivement, pour les sous-coûts $C_{\text{mfcc}}(u,v)$, $C_{\text{amp}}(u,v)$ et $C_{\text{F0}}(u,v)$.

Différentes techniques peuvent être utilisées pour estimer les poids de la fonction de coût (ALÍAS, FORMIGA et LLORÁ 2011). Notamment, même s'il existe des tentatives récentes pour construire des prédicteurs du niveau de naturel d'un système de synthèse (TAKENORI YOSHIMURA, OLIVER WATTS et K. T. u. YAMAGISHI 2016), il n'existe pas encore de mesures de qualité objectives fortement corrélées à la perception. En conséquence, la plupart des méthodes pour ajuster les poids de la fonction de coût repose sur des évaluations subjectives. Dans notre cas, comme indiqué précédemment, nous opérons à une normalisation des valeurs utilisées dans les sous-coûts.

4.1.4 Première évaluation du système

Une évaluation du système par un test de type score d'opinion moyen, *MOS*, est présentée ici. Ce test a été réalisé par 10 testeurs experts en synthèse de parole, chacun d'entre-eux a évalué 10 échantillons différents. Chaque système a donc reçu 100 notes sur une échelle de 1 à 5. A chaque étape, un seul stimuli est présenté. Ce dernier est tiré au hasard parmi des échantillons naturels, du système de synthèse ou du système en utilisant un corpus de parole réduit à une couverture de di-phonèmes.

Le système évalué ici utilise la liste de filtres présentés au paragraphe 4.1.2. La fonction de coût utilisée ici ne comporte pas de coût de sélection mais uniquement un coût de concaténation. Les voix utilisées sont les suivantes :

- *Audiobook* : il s'agit d'un corpus de voix d'homme d'environ 10h construit à partir d'un livre-audio. De part sa nature, il s'agit d'un corpus assez expressif et dont le contenu est peu contrôlé.
- *IVS* : il s'agit d'une voix d'environ 7h enregistrée spécifiquement pour les besoins de services vocaux, disposant d'une segmentation avec des corrections manuelles. En particulier, il s'agit d'une voix assez neutre particulièrement bien adaptée à la synthèse par concaténation.

Les résultats sont reportés dans le tableau 4.2. Ils permettent de montrer que le système offre une qualité de synthèse comparable à l'état de l'art.

Les éléments mis en oeuvre dans les systèmes de synthèse, en particulier les sous-coûts ou les filtres utilisés, peuvent varier dans la littérature. Par exemple, un sous-coût additionnel portant sur les différences de durées entre segments candidats peut également être utilisé. De plus, des contraintes liées à certaines dépendances contextuelles ou à la nature même des phonèmes peuvent être introduites. Dans la section suivante, nous allons étudier l'ajout d'une contrainte portant sur les caractéristiques des phonèmes à concaténer.

	Audiobook	*IVS*
Naturel	4.82 ±0.08	4.88 ±0.07
Système IRISA	3.38 ±0.25	3.17 ±0.21
Corpus bi-gram	2.14 ±0.14	1.72 ±0.08

TABLE 4.2 – *Résultats en score d'opinion moyen pour une voix d'homme et une voix de femme. L'algorithme A* est utilisé pour la sélection d'unités. Pour chaque voix, on évalue les échantillons naturels, le système proposé, ainsi que le système proposé avec un corpus réduit couvrant uniquement les diphonèmes.*

4.2 Introduction de contraintes phonologiques

Les artéfacts de concaténation apparaissent plus souvent sur certaines classes de phonèmes (Yi 1998). C'est le cas, par exemple, des phonèmes dont la réalisation dépend fortement du contexte (*i.e.* les liquides) qui peuvent montrer des variations assez importantes. Effectuer une concaténation sur de telles unités peut donc présenter des risques.

Pour contrer ce problème, différents types de contraintes peuvent être considérés pour tenter de diminuer la présence d'artéfacts. Dans la littérature, les travaux de Yi 1998 proposent la mise en place d'un système de pénalité suivant la classe phonémique du segment candidat ainsi que son contexte. Par ailleurs, Cadic, Cédric Boidin et D'Alessandro 2009 ont proposé l'ajout de contraintes sur les contextes phonologiques au moment de la construction du script d'enregistrement, préalable à l'enregistrement de la voix. Considérant cela, nous proposons ici une manière de prendre en compte une telle contrainte directement dans la fonction de coût lors de la sélection. L'idée sous-jacente est de conserver le bénéfice de cette contrainte en relâchant toutefois la nécessité de construire un script d'enregistrement dédié.

Dans cette section, le concept d'unité sandwich est présenté. Ensuite, l'intégration d'une contrainte sur le type de phonème concaténé est proposée sous la forme d'une pénalité. Une variante permettant de relaxer la pénalité dans certains cas est également présentée. Une évaluation est ensuite réalisée afin de valider cette proposition.

4.2.1 Sandwichs vocaliques

Les travaux de Yi 1998, par l'analyse de phrases contenant des artéfacts de concaténation, ont montré que des concaténations sur des voyelles ou des semi-voyelles présentent plus de risques que des concaténations sur des plosives ou encore des fricatives.

Dans (Cadic, Cédric Boidin et D'Alessandro 2009 ; Cadic et D'Alessandro 2010), la proposition est un peu différente dans le sens où les auteurs proposent un critère de couverture dont l'objectif est de maximiser le nombre d'unités sandwich dans le script d'enregistrement. Une unité sandwich est une séquence de phonèmes possédant un ou plusieurs noyaux syllabiques entourés par deux phonèmes considérés comme plus « robustes » aux concaténations. Une unité sandwich peut formellement être définie comme une unité satisfaisant l'expression rationnelle suivante :

$$R(A^*VA^*)^+R \tag{4.5}$$

où + signifie 1 ou plusieurs occurrences, * signifie 0 ou plusieurs occurrences et R, A et V sont les trois ensembles phonétiques tels que présentés dans (Cadic, Cédric Boidin et D'Alessandro 2009) :

V (voyelle) : les voyelles sont considérées comme très sensibles aux concaténations.

A (acceptable) : cet ensemble regroupe les semi-voyelles, liquides, nasales, fricatives voisées et le schwa. Ces unités sont considérées comme acceptables pour des concaténations même si elles ne sont pas sans risque.

R (résistant) : les phonèmes restants (consonnes non voisées, plosives voisées) sont considérées comme des unités sur lesquelles il faut privilégier les concaténations.

Pénalité fondée sur les classes de sandwichs

Nous choisissons ici les regroupements proposés et justifiés dans (Cadic, Cédric Boidin et D'Alessandro 2009). Bien entendu, le choix des éléments dans chaque classe est discutable, par exemple considérer que toutes les voyelles sont à exclure des concaténations n'est peut-être pas justifié. Ainsi, une première manière de favoriser les jonctions sur des phonèmes « résistants », est d'introduire une pénalité dans le coût de concaténation en fonction de la classe phonétique du segment évalué :

- 0 pour les phonèmes dans R,
- une valeur légèrement plus élevé que le plus grand coût de concaténation observé dans le corpus pour tous les phonèmes de A
- une valeur très grande pour les éléments de V. Cette valeur doit être suffisamment grande pour éviter les compensations par d'autres unités dans la séquence candidate.

Avec cette première méthode que nous noterons *sand*, le coût de concaténation C'_c devient :

$$C'_c(u, v) = C_c(u, v) + K(u, v) \tag{4.6}$$

où $K(u, v) = p(v)$ est la pénalité dépendant de la classe du phonème qui débute l'unité candidate v.

L'objectif de cette pénalité n'est pas d'agir comme un nouveau coût mais simplement d'introduire de la connaissance qui n'est pas capturée par les coûts présents. Par ailleurs, il peut être nécessaire d'adapter les classes utilisées ici en fonction de la langue.

Relâchement de la pénalité dans certains cas

Avec le système précédent, tous les phonèmes d'une même classe sont considérés et pénalisés de la même manière. Cependant, il peut y avoir des unités pour lesquelles la concaténation serait excellente même si elles appartiennent à des classes présentant plus de risques. Considérons la distribution des coûts de concaténation pour une classe, par exemple les voyelles. On peut considérer que les concaténations avec les coûts les plus faibles dans cette distribution ne doivent pas être pénalisées de la même manière que les concaténations avec les coûts les plus forts. En suivant cette idée, nous proposons d'introduire une fonction de pondération pour relaxer la pénalité suivant la place de la concaténation dans la distribution des coûts de concaténation de sa classe.

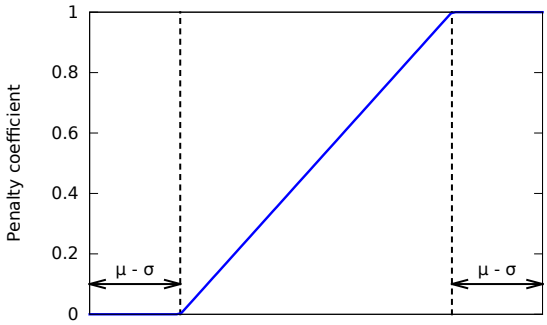

FIGURE 4.1 – *Fonction de pondération sur la distribution de coût. Le poids 0 (resp. 1) est donné aux unités dont le coût de concaténation est parmi les coûts les plus faibles (resp. les plus élevés). Entre les seuils, le poids augmente de manière linéaire.*

Cette fonction peut être définie avec le profil présenté sur la figure 4.1 et est spécifiée pour chaque sous-coût intervenant dans le coût de concaténation. Par exemple pour le F_0, on peut l'écrire de la manière suivante :

$$f_{F_0}(u,v) = \begin{cases} 0 & \text{if } C_{F_0}(u,v) < T^1_{F_0}, \\ 1 & \text{if } C_{F_0}(u,v) > T^2_{F_0}, \\ 1.0 - \frac{(T^2_{F_0} - C_{F_0}(u,v))}{(T^2_{F_0} - T^1_{F_0})} & \text{otherwise.} \end{cases} \tag{4.7}$$

Les deux seuils $T^1_{F_0}$ et $T^2_{F_0}$ utilisés ici peuvent être définis comme suit :

$$T^1_{F_0} = \mu_{C_{F_0}} - \sigma_{C_{F_0}} \tag{4.8}$$

$$T^2_{F_0} = \mu_{C_{F_0}} + \sigma_{C_{F_0}} \tag{4.9}$$

Le choix de cet intervalle de tolérance repose sur l'observation des distributions de coût. Dans l'absolu, il pourrait être optimisé pour chaque sous-coût. Au final, la pénalité associée à cette méthode est modifiée de la manière suivante :

$$K(u,v) = (f_{\text{mfcc}}(u,v) + f_{\text{amp}}(u,v) + f_{F_0}(u,v)) * p(v) \tag{4.10}$$

où $f_{\text{mfcc}}(u,v)$, $f_{\text{amp}}(u,v)$ et $f_{F_0}(u,v)$ correspondent aux fonctions de pondérations de la forme de l'équation 4.7 pour les MFCC, l'amplitude et le F_0.

Cette méthode sera notée *fuzzy-sand* par la suite. Elle permet d'appliquer une pénalité progressivement en fonction de la qualité relative de la concaténation. Comme précédemment, si une concaténation est parmi les pires de sa classe, la pénalité est appliquée entièrement. Entre les deux seuils, la pénalité augmente linéairement en fonction du rang du coût de concaténation dans la distribution.

4.2.2 Évaluation perceptive

Dans cette évaluation, les deux voix *Audiobook* et *IVS* sont utilisées. La procédure expérimentale suivie ici est celle présentée dans (Chevelu et al. 2015), en utilisant le même corpus de phrases de test. Ce dernier comporte environ 27000 phrases extraites d'un ensemble de livres de différents styles. Pour chacun des systèmes évalués (*baseline*, *sand* et *fuzzy-sand*), l'ensemble des phrases de test est synthétisé.

Pour évaluer les deux propositions, 12 tests perceptifs de type AB ont été réalisés. Ces tests se répartissent en 3 méthodes de sélection des échantillons, en utilisant 2 voix, et en comparant le système *baseline* soit au système *sand* soit au système *fuzzy-sand*. Les 3 méthodes de sélection des échantillons sont les suivantes :

Sélection Aléatoire : Les phrases sont sélectionnées au hasard parmi toutes les phrases synthétisées. Ce type de sélection permet d'évaluer, en moyenne, comment est perçue la proposition par rapport au système *baseline*.

Phrases les plus différentes : Les échantillons les plus différents sont sélectionnés. Le choix s'opère en calculant un coût d'alignement en l'échantillon généré par le système *baseline*et celui généré par le système évalué. Cette procédure est identique à celle présentée dans le chapitre 1, partie 1.2, également détaillée dans (Chevelu et al. 2015). Cette méthode permet de concentrer l'évaluation sur des échantillons qui contiennent des différences entre les systèmes, ce qui n'apparaît pas forcément lors d'une sélection aléatoire.

Phrases avec le coût de concaténation le plus élevé : les phrases qui comportent un coût de concaténation le plus élevé pour le système *baseline* sont sélectionnées. En effet, l'objectif des sandwichs est d'éviter des concaténation désastreuses. On suppose donc que si la proposition est efficace, les concaténations dans ces phrases devraient être améliorées par l'usage de la pénalité.

Chaque test a été réalisé par 10 testeurs experts, chacun évaluant 10 paires distinctes de stimuli. En conséquence, au total, 100 stimuli sont évalués par système et par test. La question posée est « Lequel des deux échantillons, selon vous, présente la meilleure qualité globale ? ». Les réponses possibles étaient alors A, B ou indifférent. Pour la dernière méthode de sélection, une seconde question, portant sur la qualité des concaténations, a également été posée.

Les tableaux 4.3 et 4.4 présentent les résultats pour les systèmes *sand* et *fuzzy-sand*. Chaque ligne correspond à un test AB. On observe que la pénalité fixe du système *sand* n'apporte pas l'amélioration attendue lorsqu'elle est intégrée directement dans la fonction de coût. En effet, pour *IVS*, les testeurs expriment en majorité des votes « indifférent », et pour *Audiobook*, ils préfèrent largement le système *baseline*. Il est à noter que cela ne remet pas en cause l'efficience des unités sandwich pour la construction de corpus mais que la transposition directe dans le moteur de synthèse n'est pas suffisante. Concernant les résultats du système *fuzzy-sand*, les choses sont très différentes puisque la préférence va nettement vers cette proposition par rapport au système *baseline*. La seule

Voix	Méthode de sélection	Question	Réponses		
			baseline	*sand*	Indifférent
IVS	Aléatoire	Qualité globale	**45%**	34%	21%
	Coût d'align.	Qualité globale	31%	34%	**35%**
	Coût de concat.	Concaténation	33%	30%	**37%**
		Qualité globale	30%	**35%**	**35%**
Audiobook	Aléatoire	Qualité globale	38%	**39%**	23%
	Coût d'align.	Qualité globale	**47%**	32%	21%
	Coût de concat.	Concaténation	**38%**	31%	31%
		Qualité globale	**39%**	30%	31%

TABLE 4.3 – *Résultats des tests AB pour le système sand. On retrouve les tests pour les deux voix en faisant varier la méthode de sélection des échantillons. La troisième colonne précise la question posée soit sur la qualité globale, soit sur la qualité des concaténations.*

Voix	Méthode de sélection	Question	Réponses		
			baseline	*sand*	Indifférent
IVS	Aléatoire	Qualité globale	35%	**40%**	25%
	Coût d'align.	Qualité globale	31%	**48%**	21%
	Coût de concat.	Concaténation	20%	**59%**	21%
		Qualité globale	27%	**49%**	24%
Audiobook	Aléatoire	Qualité globale	**43%**	42%	15%
	Coût d'align.	Qualité globale	42%	**46%**	12%
	Coût de concat.	Concaténation	33%	**38%**	29%
		Qualité globale	36%	**43%**	21%

TABLE 4.4 – *Résultats des tests AB pour le système fuzzy-sand. On retrouve les tests pour les deux voix en faisant varier la méthode de sélection des échantillons. La troisième colonne précise la question posée soit sur la qualité globale, soit sur la qualité des concaténations.*

exception est le cas de la sélection aléatoire des échantillons pour la voix *Audiobook*. Cependant, l'écart de votes ne semble pas significatif. On peut également remarquer que le nombre de votes « indifférent » diminue grandement dans le tableau 4.4. Cela traduit le fait que les différences entre le système *baseline* et le système *fuzzy-sand* sont plus facilement perceptibles.

Pour conclure cette section, l'approche *fuzzy-sand* montre un réel gain par rapport au système de base. Sa flexibilité permet d'obtenir un tri des candidats plus fin en prenant en compte un risque potentiel d'artéfact de concaténation. De manière complémentaire, une étude informelle a montré une augmentation du nombre de concaténations dans le cas *fuzzy-sand* par rapport aux deux autres systèmes. Cela montre que il ne faut pas toujours chercher à effectuer le moins de concaténations possibles, mais plutôt contrôler de manière fine où il est préférable de les effectuer.

4.3 Introduction de contraintes prosodiques

Lors de la recherche de la meilleure séquence d'unités, l'ajout de contraintes prosodiques est souhaitable, notamment pour le contrôle des durées phonémiques. Cependant, l'ajout d'un coût cible local au phonème peut amener à sélectionner une séquence de phonèmes globalement satisfaisante mais faisant apparaître localement un problème important de durée. L'idée développée ici est d'introduire un coût sur la durée des phonèmes dépendant des choix précédents dans la séquence.

Dans cette partie, le prédiction de la durée des phonèmes par réseau de neurones est tout d'abord abordée. Ensuite, la proposition de coût sur la durée est développée et enfin une évaluation perceptive est réalisée.

4.3.1 Prédiction des durées des phonèmes

La prédiction des durées des phonèmes est un problème largement étudié en synthèse de parole. Différentes approches ont été employées allant des approches par règles dérivées par des experts, jusqu'à l'usage d'arbres de décision comme dans HTS (Yoshimura et al. 1999) ou de réseaux de neurones, par exemple dans (Riedi 1995 ; Karaali et al. 1998). Encore récemment des travaux portent sur ce problème (Goubanova et Simon King 2008).

Dans ce travail, un réseau de neurones de type perceptron multi-couches est utilisé. Les attributs en entrée sont de dimension 250 et prennent en compte des caractéristiques linguistiques et phonétiques pour le phonème courant ainsi que deux phonèmes de contexte avant et après le phonème courant. Le réseau possède une couche cachée de 512 neurones avec une fonction d'activation linéaire rectifiée et 1 neurone de sortie avec une fonction d'activation linéaire.

L'erreur RMS moyenne pour la voix *IVS* est légèrement plus faible (RMS=24.24ms, std=9.07) que pour la voix *Audiobook* (RMS=26.58ms, std=6.61). Le calcul du score de Pearson montre que les prédictions du réseau de neurones sont fortement corrélées aux valeurs réelles. Une analyse détaillée des résultats par phonème montre que les phonèmes pour lesquels l'erreur RMS est la plus élevée sont ceux dont le nombre de représentants dans le corpus est faible. À titre d'exemple, seulement 2 réalisations de /ɲ/ sont présentes dans le corpus *Audiobook*. Dans la mesure où l'objectif est d'influencer la sélection, les résultats obtenus avec cette architecture sont jugés suffisants ici.

4.3.2 Proposition de coût cible pour la durée

L'objectif du coût cible proposé est d'influencer la sélection d'unités de sorte que les durées des unités sélectionnées soient, en moyenne, à la même distance des durées prédites. Une telle définition signifie qu'il est préférable d'avoir une séquence d'unités

dont la différence à la consigne est homogène plutôt que d'avoir une séquence très proche de la consigne sauf pour quelques unités où la différence est très importante.

Le coût d'une unité n dans la séquence U peut être exprimé de la manière suivante :

$$D_e = |D_t(u_n) - D(u_n)| \tag{4.11}$$

$$\Delta(u_n) = \frac{\Delta(u_{n-1}) * (n-1) + D_e}{n} \tag{4.12}$$

$$C_d(u_n) = |\Delta(u_{n-1}) - D_e| \tag{4.13}$$

avec Δ_{u_n} la distance moyenne à la durée prédite pour les unités précédentes dans la séquence (de u_1 à u_n), $D_t(u_n)$ la durée cible pour l'unité u_n, $D(u_n)$ la durée de u_n et $C_d(u_n)$ le coût local de durée pour l'unité u_n.

L'équation (4.11) calcule la différence locale de durée entre la durée prédite de l'unité cible et la durée de l'unité candidate courante. Cette différence locale D_e est utilisée pour mettre à jour la distance moyenne à la consigne $\Delta(u_n)$ dans l'équation (4.12). Le coût local de durée, $C_d(u_n)$, est ensuite calculé à partir de D_e et de $\Delta(u_n)$. Ainsi la qualité d'une unité vis-à-vis de la durée va dépendre non seulement de l'écart local à la consigne mais également de l'écart moyen observé avec les unités précédentes sur le même chemin. En d'autres termes, cela signifie que le coût associé à une unité u_n plus longue que souhaitée sera faible si les unités précédentes sont également plus longues. Le choix de la meilleure unité locale dépend ainsi des choix précédents, et des séquences comportant des unités globalement plus courtes ou plus longues seront considérées de manière équivalente. De cette manière, nous cherchons à privilégier la consistance entre les différentes unités en introduisant potentiellement un ralentissement ou une accélération globale crédible.

4.3.3 Évaluation perceptive

Afin de valider la proposition, nous avons mené une évaluation perceptive en comparant le système sans contrôle de la durée, *Uncontrolled*, au système, *Controlled*, utilisant le coût proposé dans le paragraphe précédent. Pour ce dernier, nous avons fixé de manière expérimentale les poids W_{tc} et W_{cc} au valeurs de 30 et 70 respectivement, de manière à équilibrer les contributions entre coût cible et coût de concaténation.

Deux tests de type AB ont été menés avec 13 testeurs pour le premier et 11 pour le second. L'ensemble des testeurs était constitué pour moitié d'experts. Trois réponses était proposées à chaque étape à l'évaluateur lui permettant de donner sa préférence : A, B ou Indifférent. Les deux voix ont été mélangées dans chaque test.

Le premier test comportant 20 stimuli pour chaque voix à évaluer. Les testeurs avaient pour consigne d'évaluer le rythme de la parole et de choisir l'échantillons qui leur semblait le meilleur sur ce critère. Les résultats de cette première évaluation n'ont pas permis de départager les systèmes avec 43% de préférence pour *Uncontrolled* et 38% pour *Controlled* avec des intervalles de confiance à 95% non disjoints. Ces résultats suggèrent à

première vue que les deux systèmes sont équivalents. Toutefois, une analyse plus fine des stimuli présentés nous a montré que très peu d'échantillons présentaient des problèmes de durée significatifs. Un point important est que la voix *IVS* est assez neutre et les échantillons de parole synthétisés ne présentent pas de manière fréquente des problèmes de durée. Cela est moins vrai pour *Audiobook* qui est une voix beaucoup plus expressive et de ce fait les échantillons synthétisés sont plus sujets à des problèmes de durée, parfois importants.

En conséquence, nous avons choisi d'effectuer un deuxième test perceptif en se focalisant sur les échantillons présentant des artéfacts de durée audibles. 22 échantillons comportant des problèmes de durée, plus ou moins importants, ont été sélectionnés à partir des échantillons produits par le système *Uncontrolled* (11 pour chaque voix). À chaque étape d'un test AB, chaque échantillon sélectionné a été comparé au même échantillon produit par le système *Controlled*. Les résultats de ce test sont présentés sur la figure 4.2. Cette fois-ci, une préférence pour le système *Controlled* peut être observée pour les deux voix. Cette préférence est encore plus nette dans le cas de la voix *Audiobook* qui est plus expressive.

Ces différents résultats montrent que le contrôle de la durée proposé apporte un gain lorsque des artéfacts sont présents, sans toutefois détériorer les échantillons qui ne possèdent pas de problème important. Ils permettent également d'envisager le contrôle de la mélodie sur un schéma similaire en favorisant un écart constant à une consigne plutôt qu'une optimisation locale par rapport à une cible.

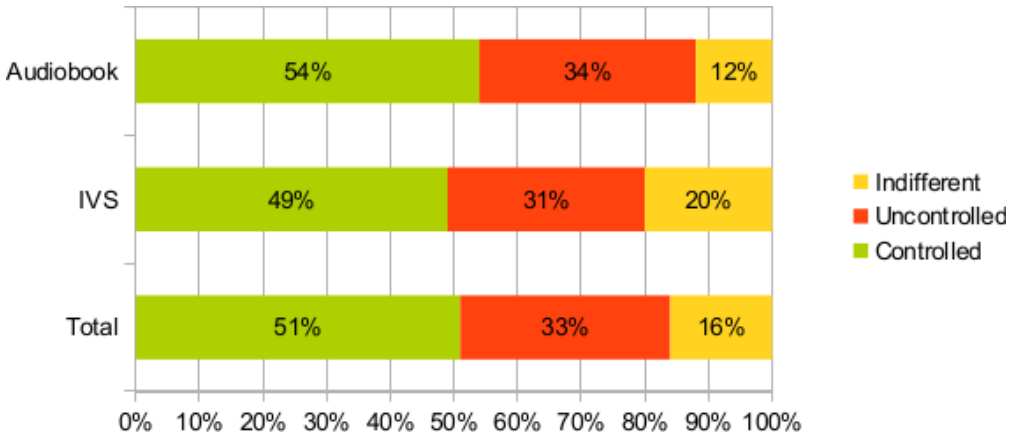

FIGURE 4.2 – *Résultats du test AB pour le coût de durée. Les systèmes Uncontrolled et Controlled sont comparés sur des phrases contenant des problèmes de durée perceptibles. L'évaluation est menée pour les deux voix IVS et Audiobook. Les résultats montrent une préférence pour le système Controlled pour lequel le coût de durée est utilisé.*

4.4 Évaluation du système avec d'autres langues : challenge Blizzard

Dans cette partie, les travaux menés dans le cadre de notre participation au challenge Blizzard sont présentés. Pour notre première participation, en 2015, la tâche principale était la synthèse de langues indiennes. En 2016 et 2017, la communauté s'est orientée vers la synthèse de livres audios adressés aux enfants, en langue anglaise.

4.4.1 Évaluation avec des langues indiennes

Contexte

En 2015, nous avons décidé de participer au challenge international de synthèse de parole Blizzard, pour la première fois. L'objectif était tout d'abord de démontrer notre capacité à adapter le système de synthèse à de nouvelles langues et également d'évaluer la qualité du système par rapport à la communauté académique sur la scène internationale.

La tâche principale du challenge était de construire un système de synthèse pour chacune des six langues suivantes : Bengali, Hindi, Malayalam, Tamil, Telugu et Marathi. Environ 4 heures de parole étaient disponibles pour trois des langues (Hindi, Tamil and Telugu), et environ 2 heures pour les trois autres (Marathi, Bengali and Malayalam). Tous les corpus de parole ont été enregistrés par des locuteurs professionnels avec une qualité studio.

Un premier problème posé par l'utilisation de langues indiennes est le changement de système d'écriture par rapport aux langues d'europe de l'ouest, et les différences d'alphabet existants entre les différentes langues indiennes manipulées. Également, la méconnaissance de ces langues amène à des difficultés afin de repérer les mots ou encore les syllabes. Un second problème est l'existence et la disponibilité d'outils, par exemple de conversion graphème vers phonèmes ou encore de segmentation automatique de la parole en phonèmes. Enfin, lors de la mise au point d'un système, il est important de disposer de locuteurs, dans l'idéal, dont la langue maternelle est celle synthétisée.

Méthodologie

Pour chaque voix, seuls le signal de parole et le texte correspondant étaient fournis. Considérant les problèmes évoqués précédemment, nous avons dû trouver des outils permettant de phonétiser un texte et segmenter un flux de parole en phonèmes. Pour cela, l'outil *eSpeak* (Duddington 2012) a été utilisé avec 4 des 6 langues : Bengali, Hindi, Malayalam, and Tamil. Les deux langues restantes n'étant pas reconnues par l'outil, nous avons recherché des solutions alternatives. Une solution a été trouvée pour le Telugu en utilisant une technique de translitération avec un script IT3. Une fois les textes

phonétisés, ou translitérés, les signaux de parole ont été segmentés en phonème en utilisant l'outil *MAUS* (Schiel 1999). Cet outil a été choisi notamment parce qu'il fournit des modèles de phonèmes génériques. Nous avons également utilisé l'outil *ROOTS* (Chevelu, Lecorvé et Lolive 2014a) afin de stocker de manière cohérente toutes les informations nécessaires et pour réaliser les conversions depuis IPA (format de sortie de l'outil *eSpeak*) vers l'alphabet SAMPA (utilisé par *MAUS*).

Le moteur de synthèse présenté dans ce chapitre a ensuite pu être utilisé avec l'algorithme de recherche A* pour optimiser la séquence d'unités. La pénalité sandwichs présentée dans la section 4.2, équation (4.10) a été employée. Par ailleurs, les filtres de pré-sélection des unités suivants ont été utilisés :

1. Label du segment associé, diphonème ou autre (ne peut être relâché).
2. Est-ce un *Non Speech Sound* (ne peut être relâché) ?
3. Est-ce un phonème nasal ?
4. Est-ce un phonème long ?
5. Est-ce un phonème accentué (accent primaire) ?
6. Est-ce un phonème accentué (accent secondaire) ?
7. Le phone est-il dans la dernière syllabe du groupe de souffle ?
8. Le phone est-il dans la dernière syllabe de la phrase ?
9. La syllabe courante est-elle en fin de mot ?

Ces filtres ont été choisis de manière à prendre en compte des caractéristiques phonétiques et prosodiques importantes pour les langues visées. De plus, n'ayant pas trouvé de locuteur parlant une de ces langues avant les derniers jours précédant la soumission de nos échantillons, nous n'avons pas pu optimiser correctement la configuration du système.

Résultats et discussion

Les évaluations perceptives menées lors du challenge reposent sur deux sous-corpus pour chaque langue : un ensemble de phrases lues (RD) et un ensemble de phrases sémantiquement imprévisibles (SUS). Trois tests ont été menés avec des testeurs rémunérés pour mesurer *la similarité au locuteur*, *le naturel*, et *l'intelligibilité*. Les deux premiers critères ont été évalués grâce à des tests *MOS* tandis que l'intelligibilité a été évaluée par un taux d'erreur mot (Word Error Rates - WER), résultat d'une retranscription des mots reconnus par le testeur. Plus de détails sur le protocole d'évaluation sont fournis dans (Prahallad, Vadapalli et al. 2015).

Les résultats des évaluations pour le système de l'IRISA sont présentés dans le tableau 4.5. Les nombres en gras indiquent que notre système a obtenu les meilleurs résultats parmi tous les participants. Les résultats en terme de similarité sont plutôt bons dans la mesure où ils figurent parmi les meilleurs dans la moitié des cas. Cela n'est

pas surprenant dans la mesure où il s'agit dans notre cas d'un système par sélection d'unités. Concernant le naturel, notre système obtient des résultats moyens. Pour toutes les langues, les résultats sont plus bas sur l'ensemble SUS, ce qui peut être expliqué par l'apparition de mots normalement peu utilisés conjointement. Cela peut entraîner l'utilisation de diphones rares et une chute dans la qualité des concaténations. Pour le Télugu, le score de similarité est très élevé (4.2) pour notre système, et assez proche du naturel (4.5) alors que le score de similarité pour le second meilleur système est bien plus bas (3.1). La seule différence pour le Télugu est que nous n'avions pas de phonétiseur, et que nous avons opéré à une translitération directe en utilisant IT3.

Un point intéressant est l'écart important entre score de similarité et score de naturel pour le Télugu sur l'ensemble SUS. Cette chute est nettement moins forte pour les autres langues. Comme mentionné précédemment, pour le Télugu, nous avons utilisé une translitération en lieu et place de la phonétisation. Cependant, vus les résultats sur l'ensemble RD, nous pensons que cela amène une qualité équivalente ici à la phonétisation. Une explication possible est que le locuteur pour le corpus de parole de Télugu est un commentateur professionnel avec un débit de parole élevé ainsi qu'une voix assez expressive. Cet effet est amplifié par la construction des phrases de l'ensemble SUS.

L'intelligibilité des phrases de l'ensemble SUS, telles que synthétisées par notre système, semble assez basse. Malgré tout, ces résultats sont globalement du même ordre que ceux des autres systèmes. La seule exception est celle du Bengali avec un taux d'erreur mot de 100% (seulement 2 systèmes ont cette valeur). Néanmoins, des points méthodologiques discutables peuvent être mis en avant afin d'expliquer ces taux d'erreur élevés, ainsi que certaines explications données par les organisateurs du challenge :

- Les locuteurs natifs ne sont pas habitués à saisir des mots avec les scripts des langues indiennes, aucun clavier standardisé n'existe
- Pour certains testeurs, il s'agissait de la première fois qu'il devait saisir des phrases complètes dans ces langues.
- Les APIs de translitération de Google nécessitent l'entrée d'un espace afin que le caractère ASCII soit changé en caractère UTF8. Cet espace est souvent oublié par les testeurs.
- Le taux d'erreur mot est calculé de manière binaire, considérant par exemple la distinction entre une voyelle courte ou longue comme une erreur.

De manière globale, les résultats obtenus sont comparables aux autres systèmes ayant participé au challenge. Ce sont des résultats satisfaisants pour une première participation au challenge avec des connaissances très limitées sur les langues employées. De plus, la quantité de données utilisées pour le challenge était assez limitée, et nous devrions logiquement obtenir de meilleurs résultats en augmentant la taille des corpus.

Langue	**Similarité**		**Naturel**		**WER**
	RD	SUS	RD	SUS	
Bengali	2.9 (1.08)	2.3 (1.14)	2.7 (1.08)	2.1 (0.93)	100 (0)
Hindi	3.5 (1.11)	3.3 (1.10)	2.8 (1.02)	3.2 (1.09)	31 (21)
Malayalam	**3.0** (1.24)	**3.2** (1.36)	2.7 (0.90)	**2.9** (0.87)	73 (18)
Tamil	**3.6** (1.11)	**3.4** (1.19)	3.2 (0.97)	3.0 (1.22)	50 (24)
Telugu	**4.2** (0.97)	1.9 (1.04)	2.9 (1.10)	2.1 (0.86)	62 (19)

TABLE 4.5 – *Résultats des tests pour la similarité et le naturel avec un score MOS sur une échelle entre 1 et 5, ainsi que l'intelligibilité en taux d'erreur mot (WER). Pour la similarité et le naturel, les tests sont effectués sur un ensemble de phrases lues (RD) et un ensemble de phrases sémantiquement imprévisibles (SUS). Pour chaque résultat, la moyenne et la déviation standard sont indiquées. Le taux d'erreur mots est quant à lui évalué sur l'ensemble SUS (en %).*

4.4.2 Évaluation en anglais avec des livres audios pour enfants

Contexte

L'objectif des éditions 2016 et 2017 du challenge était de montrer la capacité de la synthèse de parole à générer une parole expressive et de qualité. La tâche adressée est la génération de livres-audio pour les enfants en Anglais. Une description du challenge, des participants et des résultats obtenus sont présentés dans (Simon KING et KARAISKOS 2016).

La principale difficulté avec les livres-audio, en particulier ceux pour les enfants, est le changement de personnage, et ici l'imitation d'animaux (i.e. rugissement) ou l'apparition d'autres sons (i.e. cloche pour indiquer un changement de page). En considérant l'expressivité de la voix ainsi que les différents sons ou personnages que l'on peut trouver dans les livres fournis, les principaux challenges sont la segmentation en phonèmes et le contrôle de l'expressivité.

Au final, l'objectif était de construire une voix expressive de la meilleure qualité possible avec un corpus d'environ 5h de parole. Le corpus était constitué d'un ensemble de 50 livres audios fourni avec le signal et le texte correspondant.

Méthodologie

Nous avons utilisé pour ce challenge le moteur de synthèse de l'équipe configuré avec un coût cible incluant les attributs définis dans le tableau 4.1. Par ailleurs, trois filtres supplémentaires ont été utilisés pour prendre en compte une variabilité plus importante dans le corpus. Les deux premiers portent sur le contour intonatif (montant/descendant) porté par la syllabe. Le troisième indique si le segment est issu d'un dialogue ou d'une

zone de narration. Comme pour notre participation précédente, les pénalités sandwich ont été ajoutées au coût de concaténation, comme présenté dans la section 4.2, équation (4.10).

Résultats et discussion

Trois types d'évaluations ont été menées pour cette édition du challenge. La première concerne des paragraphes de livres-audios et tente d'évaluer les dimensions suivantes :

- Impression globale (« mauvais » à « excellent »),
- Agréabilité (« très déplaisant » à « très plaisant »),
- Pauses (« pauses désagréables » à « pauses appropriées, agréables »),
- Accents (« accentuation non naturelle » à « accentuation naturelle »),
- Intonation (« la mélodie ne correspond pas au type de phrase » à « la mélodie correspond au type de phrase »),
- Émotion (« aucune expression d'émotions » à « expression des émotions authentiques »),
- Effort d'écoute (« très fatiguant » à « très facile »).

Chacun de ces critères est évalué sur une échelle MOS. Cette évaluation permet de mettre la synthèse dans son contexte d'usage en évaluant des paragraphes entiers. Cependant, la longueur des paragraphes rend cette évaluation difficile. En particulier, il n'est pas nécessaire d'utiliser des échantillons aussi longs pour évaluer la similarité avec le locuteur cible ainsi que le naturel de la voix. Pour ces deux derniers critères, une seconde évaluation est conduite sur des phrases isolées. Enfin, une troisième évaluation est conduite sur des phrases de type SUS afin de mesurer l'intelligibilité. Au cours des différentes évaluations, 17 systèmes sont comparés incluant le naturel, excepté pour l'intelligibilité pour laquelle il n'y a pas de naturel.

Les résultats pour notre système sont présentés dans les tableaux 4.6 et 4.7. Pour chaque critère, notre système obtient des résultats moyens, excepté pour le placement des pauses qui obtient un score très faible. Ce résultat s'explique par le fait qu'un problème logiciel était présent dans la brique de prédiction des pauses, amenant à une incohérence entre les résultats lors de la mise au point du module de prédiction et son utilisation en phase de test. Pour les autres critères, les résultats moyens sont dûs à la qualité de la segmentation automatique en phones, à la variabilité importante des réalisations acoustiques des différents phonèmes, ainsi qu'à la configuration largement sous-optimale du système. En effet, l'adaptation d'un système de segmentation sur seulement 4h de parole, comportant beaucoup de variabilité, n'a pas permis ici d'obtenir une segmentation fiable. De plus, les éléments utilisés dans la fonction de coût cible, ainsi que les filtres, ont été choisis et pondérés de manière empirique. Une optimisation de ces choix devrait être effectuée par des évaluations systématiques pour améliorer le système.

Critère	Score	Rang	Naturel	Min	Max
Impression globale	$2,1$	$9/17$	$4,9$	$1,6$	$3,9$
Agréabilité	$2,4$	$8/17$	$4,8$	$1,6$	$3,9$
Pauses	$1,8$	$17/17$	$4,8$	$1,8$	$3,6$
Accents	$2,1$	$13/17$	$4,8$	$2,0$	$3,6$
Intonation	$2,2$	$10/17$	$4,9$	$2,0$	$3,8$
Émotion	$2,7$	$6/17$	$4,8$	$2,1$	$3,8$
Effort d'écoute	$1,8$	$14/17$	$4,9$	$1,6$	$3,8$

TABLE 4.6 – *Résultas des tests MOS pour différents critères dans le cas de paragraphes de livres audios, avec tous les testeurs. Pour chaque critère, les valeurs du score ainsi que le rang du système par rapport aux autres sont donnés. Le score du naturel, le système avec le moins bon score et le système avec le meilleur score (hors naturel) sont indiqués.*

Critère	Score	Rang	Naturel	Min	Max
Similarité	$3,6$	$5/17$	$4,7$	$1,6$	$4,2$
Natural	$2,8$	$6/17$	$4,8$	$1,9$	$4,2$

TABLE 4.7 – *Résultas des tests MOS pour la similarité et le naturel avec tous les testeurs. Pour chaque critère, les valeurs du score ainsi que le rang du système par rapport aux autres sont donnés. Le score du naturel, le système avec le moins bon score et le système avec le meilleur score (hors naturel) sont indiqués.*

Il est à noter de manière intéressante que le système s'en sort convenablement pour ce qui est de l'émotion (score de $2,7$, rang $6/17$). Pour l'« agréabilité » (score de $2,4$, rang $8/17$) et l'impression globale (score de $2,1$, rang $9/17$) de la parole générée, les scores sont moins bons mais le système reste dans la première moitié du tableau. Il est surprenant de voir la différence entre émotion, effort d'écoute et « agréabilité ». En effet, on pourrait s'attendre à ce que l'effort d'écoute soit comparable aux deux autres critères.

Pour la similarité, le système obtient de très bons résultats avec un score moyen de $3,6$ et une valeur médiane de 4. De manière similaire, le naturel de la parole générée est évalué de manière assez positive avec un score moyen de $2,8$ et une valeur médiane de 3. Ces résultats sont cohérents avec la nature du système utilisé, notamment pour la similarité. Cependant, cette évaluation semble donner de meilleurs résultats car il s'agit de phrases isolées et non de paragraphes. Les aspects prosodiques, hors du contexte du paragraphe, semblent avoir moins d'impact sur l'évaluation.

Enfin, l'intelligibilité, pour notre système, est assez faible comparé aux autres systèmes avec un taux d'erreur mot de 52%. L'explication la plus probable est la qualité moyenne de la segmentation automatique dont l'effet est décuplé par l'évaluation de l'intelligibilité sur des phrases SUS. En effet, le contexte d'apparition des mots n'aident en rien à les comprendre dans le cas présent.

4.5 Conclusion

Dans ce chapitre, nous avons présenté le système de synthèse de l'IRISA ainsi que ses différentes évolutions, notamment, à travers les travaux de thèse de David Guennec. Pour rappel, ces travaux ont porté sur différents axes : l'étude des algorithmes pour effectuer la sélection d'unités, l'amélioration du contrôle de la prosodie, et également l'amélioration du contrôle des concaténations. De plus, les deux participations de l'équipe au challenge international de synthèse de parole Blizzard ont été présentées. Les résultats obtenus à ces deux compétitions sont tout à fait corrects vis-à-vis de la communauté internationale et tout à fait encourageants.

Les perspectives d'évolution du système sont multiples afin d'améliorer la qualité de la synthèse. Un premier axe de recherche porte sur la représentation de l'espace acoustique pour la construction de fonctions de coût cible performantes et adaptables. En effet, le coût cible implique en général des choix empiriques tant sur les critères linguistiques et prosodiques à utiliser mais également sur leur pondération. Ainsi la recherche d'une représentation continue dans laquelle nous pourrions projeter les attributs décrivant les phonèmes serait une avancée. Une manière de procéder pourrait être la recherche d'embeddings de phonèmes, de manière similaire aux embeddings de mots, notamment par l'usage d'apprentissage profond.

Un deuxième axe de recherche majeur concerne la gestion de grandes bases de parole hétérogènes. En effet, des difficultés se posent dès lors que l'on cherche à utiliser de grands corpus, en termes d'efficacité, ou bien des données variées, incluant potentiellement différents types d'expressivité, d'émotions ou encore différents locuteurs. Sur cette thématique, des travaux ont débuté dans l'équipe : une thèse concernant la détection d'anormalité dans la parole et les mouvements faciaux et une thèse sur la caractérisation et la génération d'une prosodie expressive pour les livres-audio. Les travaux de la première thèse sont notamment applicables dans le cadre de grandes bases de parole assez peu contrôlées dans lesquelles des évènements acoustiques non souhaitables peuvent apparaître. Nous utilisons d'ailleurs les premiers résultats pour notre participation au challenge en 2017. La seconde thèse, qui a démarré fin 2016, va quant-à-elle adresser le problème de la prosodie expressive avec des corpus mono-locuteur de grande taille (environ 80h).

Conclusion et perspectives

Depuis 2008, ma thématique de recherche s'est fortement élargie et englobe désormais les différents niveaux d'étude nécessaires à la synthèse. Ainsi, les travaux présentés dans ce manuscrit s'inscrivent dans une démarche d'amélioration de l'expressivité en synthèse de parole. En effet, que ce soit les outils de représentation des corpus de parole, l'étude des styles de parole ou l'adaptation de la prononciation, ou encore la construction d'un moteur de synthèse, tous ces sujets contribuent à une meilleure gestion de l'expressivité. Ces travaux ont été réalisés dans le cadre de thèses que j'ai encadrées, de collaborations et de projets de recherche comme Phorevox ou SynPaFlex sur la période 2009-2016. Ils aboutissent à la création d'une chaîne complète de synthèse et à la participation au challenge Blizzard, montrant ainsi l'impact des différentes contributions. Les principales perspectives de l'ensemble de ces travaux, présentées ci-après, sont multiples et prolongent les axes de recherche développés dans ce manuscrit.

Prononciation, disfluences et parole spontanée La prolongation des travaux sur la prononciation (chapitre 2) et la prise en compte de phénomènes omniprésents en parole spontanée et/ou expressive constituent la première perspective. Celle-ci se situe dans la continuité des travaux de thèse de Raheel Qader, soutenue fin mars 2017. Notamment, la seconde partie de ses travaux a porté sur la modélisation et la génération des disfluences. Dans (Qader et al. 2017), un cadre de travail unifiant la modélisation de plusieurs types de disfluences (pauses, fillers, répétitions, révisions) est proposé. Au niveau de la synthèse de parole, connaître l'emplacement et la nature des disfluences permettrait d'améliorer grandement l'expressivité. Cependant, cela engendre des questions par rapport à l'obtention des données nécessaires à l'entraînement des modèles et également à la construction de voix de synthèse. De plus, une part importante de l'expressivité est également liée au niveau lexical. Une manière de prendre cela en compte est l'usage de techniques de reformulation ou de paraphrase. Cependant, cela peut requérir l'intégration d'une composante sémantique afin d'une part d'interpréter le sens du texte donné en entrée du système et également contrôler la validité de la reformulation, que ce soit pour conserver le sens ou le modifier.

Prosodie et style de parole Dans le chapitre 3, nous avons présenté des travaux en lien avec l'analyse de différents styles de parole, ainsi que la dérivation de règles pour l'application de contraintes liées au style dans le processus de synthèse. Ces travaux ont permis de mettre en évidence des problèmes sur le rythme tel que produit par un système de synthèse de parole. Dans le même temps, différentes études ont montré des manques au niveau du contrôle de la prosodie. Afin d'aller plus loin sur ces questions, et accroître la flexibilité des systèmes, nous proposons de nous focaliser sur l'étude de livres audios et d'investiguer les différences entre style narratif et dialogue, incluant les différences de jeu pour les divers personnages, ainsi que la manière de jouer les émotions lors de la lecture. Pour cela, grâce à la librairie ROOTS permettant de gérer de grand corpus et leurs annotations, nous construisons un corpus mono-locuteur de grande taille (>80 heures). Ce dernier comporte un ensemble de livres assez variés ainsi qu'une part significative de dialogues. Il est construit dans le cadre du projet SynPaFlex (décrit dans l'annexe B.2) et d'une thèse débutée en décembre 2016. Grâce à ce corpus, nous allons être en mesure d'analyser de manière statistique les différents facteurs prosodiques qui contribuent aux différents styles présents et ensuite construire des modèles prédictifs applicables en synthèse, notamment par l'usage d'apprentissage profond. Nous pourrons également analyser comment un locuteur reproduit les émotions en appliquant la méthodologie présentée dans (BARTKOVA, JOUVET et Elisabeth DELAIS-ROUSSARIE 2016).

Synthèse de parole avec de très grands corpus Les évolutions du domaine de la synthèse et de l'apprentissage automatique nous conduisent à modifier le paradigme de synthèse en prenant une orientation quasi exclusivement du domaine de l'apprentissage automatique. Les systèmes Wavenet de Google (OORD et al. 2016) et Deep Voice de Baidu (ARIK et al. 2017) en sont deux bons exemples. En suivant cette tendance, nous sommes également en train d'intégrer l'apprentissage profond dans notre système afin de mieux contrôler le processus de synthèse (par exemple, pour la génération de consignes prosodiques). Ce type d'approche requiert de grandes quantités de données. Comme mentionné précédemment, nous travaillons à la construction d'un grand corpus de parole en parallèle de ces travaux. Pour aller plus loin, nous avons également débuté des travaux à plus long terme visant à la définition d'une mesure de similarité entre phonèmes. L'idée suit celle des *embeddings de mots* et consiste à tenter de transposer ce type d'approche dans le but de construire un espace de représentation continu des phonèmes. Le résultat serait une meilleure capacité à traiter de très grands volumes de données, potentiellement hétérogènes. Un stage de master a démarré sur le sujet et nous espérons qu'il va déboucher sur une thèse. Grâce à sa flexibilité, le système de synthèse développé dans l'équipe servira de plateforme pour expérimenter ces nouvelles stratégies.

Langues peu dotées et/ou régionales À l'opposé des grands volumes de données, on trouve les langues peu dotées, qui par l'absence de données, posent de nombreux problèmes. En effet, l'absence de données a en général pour conséquence l'absence d'outils de traitement automatique pour ces langues. C'est le cas, par exemple, de l'Arménien

Occidental, pour lequel nous avons débuté, il y a quelques mois, une collaboration avec Anaïd Donabedian de l'Inalco. L'objectif est de construire un dictionnaire de l'Arménien pour lequel nous avons la charge de proposer un outil de phonétisation automatique, et également de la synthèse de parole afin de vocaliser les mots non enregistrés dans le dictionnaire. Un deuxième exemple est celui du Breton pour lequel le manque d'outils numériques est un frein à son développement (Ar Mogn 2015). Notamment, nous discutons avec l'association *Skol Vreizh* afin d'entamer une collaboration pour de tels outils qui s'inscriront dans le cadre de l'apprentissage de la langue en intégrant les variantes dialectales et l'intonation.

Au cours de mes travaux, les outils développés (Roots, moteur de synthèse notamment) ainsi que les résultats obtenus permettent d'envisager l'étude de ces différents axes de travail. Certains de ces travaux ont d'ailleurs déjà débutés, comme l'étude des disfluences en parole spontanée, ou encore l'étude des stratégies mises en œuvre par un locuteur pour refléter différents personnages.

Bibliographie

Articles de revues internationales à comité de lecture

Lolive, Damien, Nelly Barbot et Olivier Boëffard (2010). "B-spline model order selection with optimal MDL criterion applied to speech fundamental frequency stylisation". In : *IEEE Journal of Selected Topics in Signal Processing* 4.3, p. 571–581. doi : 10.1109/JSTSP.2010.2048236.

Articles de revues nationales à comité de lecture

Yoo, Hiyon, Elisabeth Delais-Roussarie, Damien Lolive et Nelly Barbot (2015). "Le Rythme en Lecture Oralisée (parole synthétique et parole naturelle)". In : *Revue Française de Linguistique Appliquée* XX.2, p. 63–77.

Articles de conférences internationales à comité de lecture

Barbot, Nelly, Olivier Boëffard et Damien Lolive (2005). "F0 stylisation with a free-knot b-spline model and simulated-annealing optimization". In : *Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech)*. Lisboa, Portugal.

Lolive, Damien, Nelly Barbot et Olivier Boëffard (2006a). "Comparing b-spline and spline models for f0 modelling". In : *Lecture Notes in Artificial Intelligence - Proceedings of the 9th International Conference on Text, Speech and Dialogue*. Brno, Czech Republic.

Lolive, Damien, Nelly Barbot et Olivier Boëffard (2006b). "Melodic contour estimation with b-spline models using a MDL criterion". In : *Proceedings of the 11th International Conference on Speech and Computer (SPECOM)*. Saint Petersburg, Russia.

Lolive, Damien, Nelly Barbot et Olivier Boëffard (2007a). "Clustering algorithm for f0 curves based on hidden markov models". In : *Proceedings of the 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*. Bonn, Germany.

Lolive, Damien, Nelly Barbot et Olivier Boëffard (2007b). "Unsupervised HMM classification of f0 curves". In : *Interspeech*. Antwerp, Belgium.

Lolive, Damien, Nelly Barbot et Olivier Boëffard (2008a). "Pitch and duration transformation with non parallel data". In : *4th conference of Speech Prosody*. Campinas, Brazil, p. 111–114.

Lolive, Damien, Nelly Barbot et Olivier Boëffard (2009). "An evaluation methodology for prosody transformation systems based on chirp signals". In : *Interspeech*. Brighton, United Kingdom, p. 2635–2638.

Barbot, Nelly, Vincent Barreaud, Olivier Boëffard, Laure Charonnat, Arnaud Delhay, Sébastien Le Maguer et Damien Lolive (2011). "Towards a Versatile Multi-Layered Description of Speech Corpora Using Algebraic Relations". In : *Conference of the International Speech Communication Association (Interspeech)*. Florence, Italy, p. 1501–1504.

Boëffard, Olivier, Charonnat Laure, Sébastien Le Maguer et Damien Lolive (2012). "Towards Fully Automatic Annotation of Audio Books for TTS". In : *LREC - Eighth International Conference on Language Resources and Evaluation*. Istanbul, Turkey.

Avanzi, Mathieu, George Christodoulides, Elisabeth Delais-Roussarie, Nelly Barbot et Damien Lolive (2014). "Towards the Adaptation of Prosodic Models for Expressive Text-To-Speech Synthesis". In : *Interspeech*. ISCA. Singapore, Singapore.

Chevelu, Jonathan, Gwénolé Lecorvé et Damien Lolive (2014a). "ROOTS : a toolkit for easy, fast and consistent processing of large sequential annotated data collections". In : *Language Resources and Evaluation Conference (LREC)*. Reykjavik, Iceland.

Delais-Roussarie, Elisabeth, Damien Lolive, Hiyon Yoo, Nelly Barbot et Olivier Rosec (2014). "Adapting prosodic chunking algorithm and synthesis system to specific style". In : *Interspeech*. ISCA. Singapore, Singapore.

Guennec, David et Damien Lolive (2014a). "Unit Selection Cost Function Exploration Using an A* based Text-to-Speech System". In : *International Conference on Text, Speech and Dialogue (TSD)*. Brno, Czech Republic.

Le Maguer, Sébastien, Elisabeth Delais-Roussarie, Nelly Barbot, Mathieu Avanzi, Olivier Rosec et Damien Lolive (2014b). "Prosodic chunking algorithm for dictation with the use of speech synthesis". In : *Proc. of Speech Prosody*. Dublin, Ireland.

Alain, Pierre, Jonathan Chevelu, David Guennec, Gwénolé Lecorvé et Damien Lolive (2015). "The IRISA Text-To-Speech System for the Blizzard Challenge 2015". In : *Blizzard Challenge 2015 Workshop*. Berlin, Germany, 4 p., 2 columns.

Chevelu, Jonathan et Damien Lolive (2015). "Do not build your TTS training corpus randomly". In : *Proceedings of the European Signal Processing Conference (EUSIPCO)*. Nice, France.

Chevelu, Jonathan, Damien Lolive, Sébastien Le Maguer et David Guennec (2015). "How to Compare TTS Systems : A New Subjective Evaluation Methodology Focused on Differences". In : *Interspeech*. Dresden, Germany.

Guennec, David, Jonathan Chevelu et Damien Lolive (2015). "Defining a Global Adaptive Duration Target Cost for Unit Selection Speech Synthesis". In : *International Conference on Text, Speech and Dialogue (TSD)*. Proceedings of International Conference on Text, Speech and Dialogue (TSD). PLZEŇ, Czech Republic, p. 157–165.

Lecorvé, Gwénolé et Damien Lolive (2015). "Adaptive Statistical Utterance Phonetization for French". In : *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brisbane, Australia, 5 p., 2 columns.

Qader, Raheel, Gwénolé Lecorvé, Damien Lolive et Pascale Sébillot (2015). "Probabilistic Speaker Pronunciation Adaptation for Spontaneous Speech Synthesis Using Linguistic Features". In : *International Conference on Statistical Language and Speech Processing (SLSP)*. Budapest, Hungary, p. 229–241.

Alain, Pierre, Jonathan Chevelu, David Guennec, Gwénolé Lecorvé et Damien Lolive (2016). "The IRISA Text-To-Speech System for the Blizzard Challenge 2016". In : *Blizzard Challenge 2016 workshop*. Cupertino, United States.

Fayet, Cédric, Arnaud Delhay, Damien Lolive et Pierre-François Marteau (2016a). "Big Five vs. Prosodic Features as Cues to Detect Abnormality in SSPNET-Personality Corpus". In : *Interspeech*. Stockholm, Sweden.

Fayet, Cédric, Arnaud Delhay, Damien Lolive et Pierre-François Marteau (2016b). "First Experiments to Detect Anomaly Using Personality Traits vs. Prosodic Features". In : *Interspeech*. London, United Kingdom.

Guennec, David et Damien Lolive (2016a). "On the suitability of vocalic sandwiches in a corpus-based TTS engine". In : *Interspeech*. San Francisco, United States.

Tahon, Marie, Raheel Qader, Gwénolé Lecorvé et Damien Lolive (2016a). "Improving TTS with corpus-specific pronunciation adaptation". In : *Interspeech*. San Francisco, United States.

Tahon, Marie, Raheel Qader, Gwénolé Lecorvé et Damien Lolive (2016b). "Optimal feature set and minimal training size for pronunciation adaptation in TTS". In : *International Conference on Statistical Language and Speech Processing (SLSP)*. Pilsen, Czech Republic.

Articles de conférences nationales à comité de lecture

Lolive, Damien, Nelly Barbot et Olivier Boëffard (2006c). "Modélisation b-spline de contours mélodiques avec estimation du nombre de paramètres libres par un critère MDL". In : *Actes des XXVIèmes Journées d'Etudes sur la Parole*. Dinard, France.

Lolive, Damien, Nelly Barbot et Olivier Boëffard (2006d). "Proposition d'un critère MDL pour l'estimation de courbes ouvertes modélisées par des b-splines". In : *Actes de la 8ème Conférence Francophone sur l'Apprentissage Automatique*. Trégastel, France.

Lolive, Damien, Nelly Barbot et Olivier Boëffard (2008b). "Transformation de la prosodie par adaptation MLLR de GMM". In : *Actes des XXVIIèmes Journées d'Etudes sur la Parole*. Avignon, France.

Boëffard, Olivier, Laure Charonnat, Sébastien Le Maguer, Damien Lolive et Gaëlle Vidal (2012). "Vers une annotation automatique de corpus audio pour la synthèse de parole". In : *JEP-TALN-RECITAL - conférence conjointe*. Grenoble, France, p. 731–738.

Chevelu, Jonathan, Gwénolé Lecorvé et Damien Lolive (2014b). "ROOTS : un outil pour manipuler facilement, efficacement et avec cohérence des corpus annotés de séquences". In : *Journées d'Etude sur la Parole (JEP)*. Le Mans, France.

Guennec, David et Damien Lolive (2014b). "Utilisation d'un algorithme A* pour l'analyse de la sélection d'unité en synthèse de la parole". In : *JEP - 30ème édition des Journées d'Etudes sur la Parole*. Le Mans, France.

Le Maguer, Sébastien, Elisabeth Delais-Roussarie, Nelly Barbot, Mathieu Avanzi, Olivier Rosec et Damien Lolive (2014a). "Algorithme de découpages en groupes prosodiques pour la dictée par l'usage de synthèse vocale". In : *Journées d'études sur la parole*. Le Mans, France.

Yoo, Hiyon, Sébastien Le Maguer, Elisabeth Delais-Roussarie, Nelly Barbot et Damien Lolive (2014). "Evaluation d'un algorithme de chunking appliqué à la dictée". In : *JEP - 30ème édition des Journées d'Etudes sur la Parole*. Le Mans, France.

Chevelu, Jonathan, Damien Lolive, Sébastien Le Maguer et David Guennec (2016). "Se concentrer sur les différences : une méthode d'évaluation subjective efficace pour la comparaison de systèmes de synthèse". In : *Journées d'Études sur la Parole*. Paris, France.

Delais-Roussarie, Elisabeth, Damien Lolive, Hiyon Yoo et David Guennec (2016a). "Patrons Rythmiques et Genres Littéraires en Synthèse de la Parole". In : *Journées d'Études sur la Parole*. Paris, France.

Guennec, David et Damien Lolive (2016b). "Une pénalité floue fondée phonologiquement pour améliorer la Sélection d'Unité". In : *Journées d'Études sur la Parole*. Paris, France.

Lecorvé, Gwénolé et Damien Lolive (2016). "Phonétisation statistique adaptable d'énoncés pour le français". In : *Journées d'Études sur la Parole*. Paris, France.

Möbius, Bernd, Sébastien Le Maguer, Ingmar Steiner et Damien Lolive (2016). "De l'utilisation de descripteurs issus de la linguistique computationnelle dans le cadre de la synthèse par HMM". In : *Journées d'Études sur la Parole*. Paris, France.

Qader, Raheel, Gwénolé Lecorvé, Damien Lolive et Pascale Sébillot (2016). "Adaptation de la prononciation pour la synthèse de la parole spontanée en utilisant des informations linguistiques". In : *Journées d'Études sur la Parole*. Paris, France.

Qader, Raheel, Gwénolé Lecorvé, Damien Lolive et Pascale Sébillot (2017). "Ajout automatique de disfluences pour la synthèse de la parole spontanée : formalisation et preuve de concept". In : *Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*. Orléans, France.

Rapports de recherche

Qader, Raheel, Gwénolé Lecorvé, Damien Lolive et Pascale Sébillot (2014). *Phonology Modelling for Expressive Speech Synthesis : a Review*. Research Report PI-2020. IRISA, équipe EXPRESSION, 18 p., 1 column.

Manuscript de thèse

Lolive, Damien (2008). "Prosody transformation : application to speech synthesis and voice transformation". Theses. Université de Rennes 1.

Autres références

Saussure de, Ferdinand (1916). *Cours de linguistique générale*. édition originale : 1916, édition 1979 : Payot, Paris. isbn : (ISBN 2-2285-0068-2).

Lindblom, B (1963). "Spectrographic study of vowel reduction". In : *The Journal of the Acoustical Society of America* 35.November 1963, p. 1773–1781. issn : 00014966. doi : 10.1121/1.1918816.

Viterbi, a.J. (1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". In : *IEEE Transactions on Information Theory* 13.2, p. 260–269. issn : 00189448. doi : 10.1109/TIT.1967.1054010.

Hart, PE, NJ Nilsson et B. Raphael (1968). "A formal basis for the heuristic determination of minimum cost paths". In : *Systems Science and ...*

Sakoe, H. et S. Chiba (1978). "Dynamic programming algorithm optimization for spoken word recognition". In : *IEEE Transactions on Acoustics, Speech and Signal Processing* 26.1, p. 43–49.

Fónagy, Ivan (1980). "L'accent français : accent probabilitaire (dynamique d'un changement prosodique)". In : *Studia Phonetica Montréal* 15, p. 123–233.

Nilsson, Nils J. (1982). *Principles of Artificial Intelligence*. Springer-Verlag. isbn : 3-540-11340-1.

Verluyten, Paul (1982). "Recherches sur la prosodie et la métrique du français". Thèse de doct. Universitaire Instelling Antwerpen.

Cahn, J. E. (1990). "The Generation of Affect in Synthesized Speech". In : *Journal of American Voice I/O Society* 8, p. 1–19.

Moulines, Eric et F. Charpentier (1990). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones". eng. In : *Speech communication* 9.5-6, p. 453–467. issn : 0167-6393.

Carlson, R. (1992). "Quarterly Progress and Status Report Synthesis : modeling variability and constraints". In : *Speech Communication* 11.2-3, p. 159–166.

Black, Alan W et Paul Taylor (1994). “CHATR : a generic speech synthesis system”. In : *15th conference on Computational linguistics*. Association for Computational Linguistics, p. 983–986.

Catach, N., C. Gruaz et D. Duprez (1995). *L'orthographe française*. Editions Nathan.

Riedi, Marcel (1995). “A neural-network-based model of segmental duration for speech synthesis.” In : *in Proceedings of Eurospeech*.

Delais-Roussarie, Elisabeth (1996). “Phonological phrasing and accentuation in French”. In : *Dam Phonology : HIL phonology papers II, den Haag : Holland Academic Graphics*. Sous la dir. de M. Nespor et N. Smith, p. 1–38.

Hunt, A. J. et Alan W. Black (1996). “Unit selection in a concatenative speech synthesis system using a large speech database”. In : *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, p. 373–376.

Hunt, Andrew J. et Alan W. Black (1996). “Unit selection in a concatenative speech synthesis system using a large speech database”. In : *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. T. 1. IEEE, p. 373–376.

Murray, I. R. et J. L. Arnott (1996). “Synthesizing emotions in speech : is it time to get excited ?” In : *Proc. of ICSLP*. vol. 3, p. 1816–1819.

Breen, A.P. et P Jackson (1998). “Non-uniform unit selection and the similarity metric within BT's Laureate TTS system”. In : *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*.

Karaali, Orhan, Gerald Corrigan, Ira Gerson et Noel Massey (1998). “Text-to-speech conversion with neural networks : a recurrent TDNN approach”. In : *in Proceedings of Interspeech*.

Taylor, Paul, Alan W Black et Richard Caley (1998). “The architecture of the Festival speech synthesis system”. In : *Proc. of the ESCA Workshop in Speech Synthesis*, p. 147–151.

Yi, J.R.W. (1998). *Natural-sounding speech synthesis using variable-length units*. Rapp. tech. Massachusetts Institute of Technology.

Di Cristo, Albert (1999). “Le cadre accentuel du français contemporain : essai de modélisation. Première partie”. In : *Langues* 2.3, p. 184–205.

Fosler-Lussier, Eric et al. (1999). “Multi-level decision trees for static and dynamic pronunciation models.” In : *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*.

Schiel, Florian (1999). “Automatic phonetic transcription of non-prompted speech”. In : *Proceedings of the International Congresses of Phonetic Sciences*, p. 607–610.

Yoshimura, Takayoshi, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi et Tadashi Kitamura (1999). “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis”. In : *in Proceedings of Eurospeech*, p. 2347–2350.

Carrier, J.-P. (2000). *L'école et le multimédia*. Centre National de Documentation Pédagogique, Hachette éducation, Paris.

Conkie, Alistair, Mark C Beutnagel, Ann K Syrdal et Philip E Brown (2000). “Preselection of candidate units in a unit selection-based text-to-speech synthesis sys-

tem". In : *International Conference on Spoken Language Processing - ICSLP*. T. 3. Icslp, p. 314–317.

Post, Brechtje (2000). "Tonal and phrasal structures in French intonation". In : *The Hague : Holland Academic Graphics*.

Sjölander, K. et J. Beskow (2000). "Wavesurfer - an open source speech tool". In : *Proceedings of Interspeech*, p. 464–467.

Toutanova, Kristina et Christopher D Manning (2000). "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger". In : *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora : held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*. Association for Computational Linguistics, p. 63–70.

Barras, C., E. Geoffrois, Z. Wu et M. Liberman (2001). "Transcriber : development and use of a tool for assisting speech corpora production". In : *Speech Communication* 33.1-2, p. 5–22.

Donovan, R. E. (2001). "A new distance measure for costing spectral discontinuities in concatenative speech synthesizers". In : *ITRW*.

Lafferty, John, Andrew McCallum et Fernando CN Pereira (2001). "Conditional random fields : probabilistic models for segmenting and labeling sequence data". In :

Schröder, M. (2001). "Emotional Speech Synthesis : A Review". In : *Proc of Eurospeech*, p. 561–564.

Stylianou, Yannis et Ann K. Syrdal (2001). "Perceptual and objective detection of discontinuities in concatenative speech synthesis". In : *International Conference on Acoustics, Speech, and Signal Processing*. T. 2, p. 837–840.

Taylor, Paul, Alan W. Black et Richard Caley (2001). "Heterogeneous relation graphs as a formalism for representing linguistic information". In : *Speech Communication* 33.1-2, p. 153–174.

Bates, Rebecca et Mari Ostendorf (2002). "Modeling pronunciation variation in conversational speech using prosody". In : *ISCA Tutorial and Research Workshop (ITRW) on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*.

Black, Alan W., Paul Taylor, Richard Caley et Rob Clark (2002). *The Festival speech synthesis system*. Rapp. tech. University of Edinburgh.

Boersma, Paul (2002). "Praat, a system for doing phonetics by computer". In : *Glot international* 5.9/10, p. 341–345.

Cunningham, H., D. Maynard, K. Bontcheva et V. Tablan (2002). "GATE : an architecture for development of robust HLT applications". In : *Proceedings of the Annual Meeting of the ACL*, p. 168–175.

Bell, Alan, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory et Daniel Gildea (2003). "Effects of disfluencies, predictability, and utterance position on word form variation in English conversation". In : *The Journal of the Acoustical Society of America* 113.2.

Guyon, Isabelle et André Elisseeff (2003). "An introduction to variable and feature selection". In : *Journal of machine learning research* 3.Mar, p. 1157–1182.

Moore, Roger K (2003). "A comparison of the data requirements of automatic speech recognition systems and human listeners." In : *INTERSPEECH*.

Chen, Ken et Mark Hasegawa-Johnson (2004). "Modeling pronunciation variation using artificial neural networks for English spontaneous speech." In : *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*.

Ferrucci, D. et A. Lally (2004). "UIMA : an architectural approach to unstructured information processing in the corporate research environment". In : *Natural Language Engineering* 10.3-4, p. 327–348.

Kumar, Rohit (2004). "A genetic algorithm for unit selection based speech synthesis". In : *Eighth International Conference on Spoken Language Processing*.

Adda-Decker, Martine, Philippe Boula de Mareüil, Gilles Adda et Lori Lamel (2005). "Investigating syllabic structures and their variation in spontaneous French". In : *Speech Communication* 46.2.

Carletta, Jean, Stefan Evert, Ulrich Heid et Jonathan Kilgour (2005). "The NITE XML Toolkit : Data Model and Query Language". In : *Language Resources and Evaluation* 39.4, p. 313–334.

Marty, Nicole (2005). *Informatique et nouvelles pratiques d'écriture*. Editions Nathan.

Pitt, Mark A, Keith Johnson, Elizabeth Hume, Scott Kiesling et William Raymond (2005). "The Buckeye corpus of conversational speech : labeling conventions and a test of transcriber reliability". In : *Speech Communication* 45.1.

Scherer, K. R. (2005). "What are emotions ? And how can they be measured ?" In : *Soc. Sci. Inf.* 44.4, p. 695–729.

Viswanathan, Mahesh et Madhubalan Viswanathan (2005). "Measuring speech quality for text-to-speech systems : development and assessment of a modified mean opinion score (MOS) scale". In : *Computer Speech & Language*.

Ferrucci, D., A. Lally, D. Gruhl, E. Epstein, M. Schor, J. W. Murdock, A. Frenkiel, E. W. Brown, T. Hampp, Y. Doganata et al. (2006). "Towards an interoperability standard for text and multi-modal analytics". In : *IBM Research Report*.

Garcia, Marie-neige, Christophe D'Alessandro, Gérard Bailly, Philippe Boula De Mareüil et Michel Morel (2006). "A joint prosody evaluation of French text-to-speech synthesis systems". In : *Proc. of LREC*.

Prahallad, Kishore, Alan W Black et Ravishankhar Mosur (2006). "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis". In : *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. T. 1.

Clark, Robert AJ, Korin Richmond et Simon King (2007). "Multisyn : Open-domain unit selection for the Festival speech synthesis system". In : *Speech Communication* 49.4, p. 317–330.

Lambert, Tanya, Norbert Braunschweiler et Sabine Buchholz (2007). "How (not) to select your voice corpus : Random selection vs. phonologically balanced". In : *Proc. of SSW6*.

Zen, Heiga, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W Black et Keiichi Tokuda (2007). "The HMM-based speech synthesis system (HTS) version 2.0". In : *Speech Synthesis Workshop (SSW)*, p. 294–299.

Chevelu, Jonathan, Nelly Barbot, Olivier Boeffard et Arnaud Delhay (2008). "Comparing Set-Covering Strategies for Optimal Corpus Design." In : *LREC*.

Goubanova, Olga et Simon King (2008). "Bayesian networks for phone duration prediction". In : *Speech communication* 50.4, p. 301–311.

Yamagishi, Junichi, Zhenhua Ling et Simon King (2008). "Robustness of HMM-based speech synthesis". In : *Science And Technology*, p. 2–5.

Bell, Alan, Jason M Brenier, Michelle Gregory, Cynthia Girand et Dan Jurafsky (2009). "Predictability effects on durations of content and function words in conversational English". In : *Journal of Memory and Language* 60.1.

Cadic, Didier, Cédric Boidin et Christophe D'Alessandro (2009). "Vocalic sandwich, a unit designed for unit selection TTS". In : *Tenth Annual Conference of the International Speech Communication Association*. 1, p. 2079–2082.

Eskenazi, M. (2009). "An overview of spoken language technology for education". In : *Speech Communication* 51.10, p. 832–844.

Eskenazi, Maxine (2009). "An overview of spoken language technology for education". In : *Speech Communication* 51.10, p. 832–844.

Goldman, Jean-Philippe, Antoine Auchlin et Anne Catherine Simon (2009). "Discrimination de styles de parole par analyse prosodique semi-automatique". In : *Interface Discours Prosodie (IDP)*.

Handley, Z. (2009). "Is text-to-speech synthesis ready for use in computer-assisted language learning ?" In : *Speech Communication* 51.10, p. 906–919.

Handley, Zöe (2009). "Is text-to-speech synthesis ready for use in computer-assisted language learning ?" In : *Speech Communication* 51.10, p. 906–919.

Rebordao, Antonio Rui Ferreira, Shaikh Mostafa Al Masum, Keikichi Hirose et Nobuaki Minematsu (2009). "How to Improve TTS Systems for Emotional Expressivity". In : *in Proceedings of Interspeech*, p. 524–527.

Schröder, Marc (2009). "Expressive speech synthesis : Past, present, and possible futures". In : *Affective information processing*. Springer, p. 111–126.

Vazirnezhad, Bahram, Farshad Almasganj et Seyed Mohammad Ahadi (2009). "Hybrid statistical pronunciation models designed to be trained by a medium-size corpus". In : *Computer Speech & Language* 23.1.

Young, S. J. et al. (2009). *The HTK Book, version 3.4*. Cambridge University Engineering Department.

Cadic, Didier, Cedric Boidin et Christophe D' Alessandro (2010). "Towards optimal TTS corpora". In : *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valetta, Malta*, p. 99–104. isbn : 2-9517408-6-7.

Cadic, Didier et Christophe D'Alessandro (2010). "High Quality TTS Voices Within One Day". In : *Seventh ISCA Workshop on Speech Synthesis*.

Calhoun, Sasha, Jean Carletta, Jason M Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman et David Beaver (2010). "The NXT-format Switchboard Corpus : a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue". In : *Language Resources and Evaluation* 44.4, p. 387–419.

Gelan, Anouk (2010). "Language and Text-to-Speech technologies for highly accessible language & culture learning". In : International Association of Online Engineering.

Lavergne, Thomas, Olivier Cappé et François Yvon (2010). "Practical very large scale CRFs". In : *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Alías, Francesc, Lluís Formiga et Xavier Llorá (2011). "Efficient and reliable perceptual weight tuning for unit-selection text-to-speech synthesis based on active interactive genetic algorithms : A proof-of-concept". In : *Speech Communication* 53.5, p. 786–800. issn : 01676393. doi : 10.1016/j.specom.2011.01.004.

Avanzi, Mathieu, Nicolas Obin, Anne Lacheret et Bernard Victorri (2011). "Toward a continuous modeling of french prosodic structure : Using acoustic features to predict prominence location and prominence degree". In : *Interspeech*, p. 2033–2036.

Hinterleitner, Florian, Georgina Neitzel, Sebastian Moller et Christoph Norrenbrock (2011). "An Evaluation Protocol for the Subjective Assessment of Text-to-Speech in Audiobook Reading Tasks". In : *Proc. of Blizzard Challenge Workshop*.

Illina, Irina, Dominique Fohr et Denis Jouvet (2011). "Grapheme-to-Phoneme Conversion using Conditional Random Fields". In : *Proc. of Interspeech*, p. 2313–2316.

Obin, Nicolas (2011). "MeLos : Analysis and modelling of speech prosody and speaking style". Thèse de doct. Université Pierre et Marie Curie-Paris VI.

Post, Brechtje (2011). "The multi-faceted relation between phrasing and intonation contours in French". In : *Intonational Phrasing in Romance and Germanic : Cross-linguistic and bilingual studies*. Sous la dir. de C. Gabriel et C. Lleò. Amsterdam : Benjamins, p. 44–74.

Stolcke, Andreas, Jing Zheng, Wen Wang et Victor Abrash (2011). "SRILM at sixteen : Update and outlook". In : *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, p. 5.

Wang, Dong et S. King (2011). "Letter-to-Sound Pronunciation Prediction Using Conditional Random Fields". In : *IEEE Signal Processing Letters* 18.2, p. 122–125.

Duddington, J. (2012). *eSpeak text to speech*.

King, Simon et Vasilis Karaiskos (2012). "The blizzard challenge 2012". In : *Proc. Blizzard Challenge workshop*.

Vaudable, C. (2012). *Analyse et reconnaissance des émotions lors de conversations de centres d'appels*. Université Paris Sud - Paris XI.

Buchholz, Sabine, Javier Latorre et Kayoko Yanagisawa (2013). "Crowdsourced assessment of speech synthesis". In : *Crowdsourcing for Speech Processing*.

Dilts, Philip C (2013). "Modelling phonetic reduction in a corpus of spoken English using random forests and mixed-effects regression". Thèse de doct. University of Alberta.

Karanasou, Penny, François Yvon, Thomas Lavergne et Lori Lamel (2013). "Discriminative training of a phoneme confusion model for a dynamic lexicon in ASR". In : *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*.

Brognaux, Sandrine, Benjamin Picart, Thomas Drugman et D Louvain (2014). "Speech synthesis in various communicative situations : impact of pronunciation variations." In : *INTERSPEECH*, p. 1524–1528.

Kolluru, BalaKrishna, Vincent Wan, Javier Latorre, Kayoko Yanagisawa et Mark J. F. Gales (2014). "Generating multiple-accent pronunciations for TTS using joint sequence model interpolation". In : *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*.

Latorre, Javier, Kayoko Yanagisawa, Vincent Wan, BalaKrishna Kolluru et Mark JF Gales (2014). "Speech intonation for TTS : Study on evaluation methodology". In : *Proc. of Interspeech*.

Sainz, Iñaki, Eva Navas, Inma Hernaez, Antonio Bonafonte et Francisco Campillo (2014). "TTS evaluation campaign with a common Spanish database". In : *Proc. of LREC*.

Tihelka, Daniel, Jindřich Matoušek et Zdeněk Hanzlíček (2014). "Modelling F0 Dynamics in Unit Selection Based Speech Synthesis". In : *Text, Speech and Dialogue* 1.Springer, p. 457–464.

Ar Mogn, Olier (2015). "Langue bretonne et nouvelles technologies : une vitalité à soutenir". In : *Coloque sur les technologies pour les langues régionales de France*, p. 71–76.

Delais-Roussarie, E, B Post, M Avanzi, C Buthke, A Di Cristo, I Feldhausen, SA Jun, P Martin, T Meisenburg, A Rialland et al. (2015). "Intonational Phonology of French : Developing a ToBI system for French". In : *Intonation in Romance*. Sous la dir. de S. Frota et P. Prieto. Oxford University Press, p. 63–100.

Prahallad, Kishore, Anandaswarup Vadapalli et al. (2015). ""The Blizzard Challenge 2015"". In : *Blizzard Challenge 2015 workshop*. Berlin, Germany.

Bartkova, Katarina, Denis Jouvet et Elisabeth Delais-Roussarie (2016). "Prosodic Parameters and Prosodic Structures of French Emotional Data". In : *Speech Prosody 2016*. Speech Prosody 2016. Boston, United States.

Delais-Roussarie, Elisabeth, Damien Lolive, Hiyon Yoo et David Guennec (2016b). "Rhythmic Patterns and Literary Genres in Synthesized Speech". In : *Speech Prosody*. Boston, United States.

King, Simon et Vasilis Karaiskos (2016). ""The Blizzard Challenge 2016"". In : *Blizzard Challenge 2016 workshop*. Cupertino, United States.

Livescu, Karen, Preethi Jyothi et Eric Fosler-Lussier (2016). "Articulatory feature-based pronunciation modeling". In : *Computer Speech & Language* 36, p. 212–232.

Oord, Aäron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior et Koray Kavukcuoglu (2016). "WaveNet : A Generative Model for Raw Audio". In : *9th ISCA Speech Synthesis Workshop*, p. 125–125.

Takenori Yoshimura, Gustav Eje Henter, Mirjam Wester Oliver Watts et Keiichi Tokuda unichi Yamagishi (2016). "A Hierarchical Predictor of Synthetic Speech Naturalness Using Neural Networks". In : *in Proceedings of Interspeech*. INTERSPEECH 2016.

Arik, Sercan O, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Jonathan Raiman, Shubho Sengupta et al. (2017). "Deep Voice : Real-time Neural Text-to-Speech". In : *arXiv preprint arXiv :1702.07825*.

Annexe A

Curriculum Vitæ

État civil

Né le 08 juillet 1982, 35 ans, marié, 2 enfants

Contact

IRISA - ENSSAT
6 rue de Kerampont
CS 80518
22305 LANNION Cedex

Tel : +33 96 46 91 65
Fax : +33 96 37 01 99
damien.lolive@irisa.fr

Formation

- **2008 - Doctorat en Informatique**, IRISA - Université de Rennes I (22)
- **2005 - Master Recherche en Informatique**, ENSSAT Lannion - Université de Rennes I (22)
- **2005 - Diplôme d'ingénieur en Informatique**, Spécialité Logiciels & Systèmes Informatiques, ENSSAT de Lannion (22)
- **2002 - D.U.T. Informatique**, IUT de Limoges (87)

Parcours professionnel

- **Sept. 2016 - Août 2017 : Délégation CNRS à mi-temps**

- **Depuis Sept. 2009 - Maître de conférences en Informatique / Chercheur à l'IRISA :** J'enseigne actuellement à l'Enssat de Lannion dans les filières *Logiciels et Systèmes Informatique (LSI)* ainsi que *Informatique, Multimédia et Réseaux (IMR)*. Pour la partie recherche, je l'effectue au sein de l'Irisa dans l'équipe Expression sur la thématique de la synthèse de parole expressive.
- **2008 – 2009 - ATER en Informatique - IRISA/ENSSAT :** Au cours de cette année, j'ai eu l'occasion de renforcer la partie enseignement et de poursuivre mes recherches en me focalisant sur les méthodes d'évaluation pour la transformation de la prosodie.
- **2005 – 2008 - Doctorat en Informatique - IRISA/ENSSAT :** Mon doctorat est intitulé : *Transformation de l'intonation : application à la synthèse de parole et à la transformation de voix*. Il traite de la modélisation de la prosodie dans une perspective de synthèse de parole et de transformation de la voix.

Recherche

Le cadre de mes travaux est celui du traitement automatique de la parole avec pour application la synthèse de parole expressive. Dans ce cadre, les thématiques que j'aborde sont :

- la constitution de corpus en vue de la création de voix artificielles : cette dernière repose sur des corpus textuels annotés dont la création conditionne la qualité de la voix créée, cette thématique est donc une fondation nécessaire pour aborder les suivantes ;
- la caractérisation de l'expressivité sur un plan prosodique : il s'agit de comprendre ce qui fait l'expressivité d'une voix, de manière à pouvoir la reproduire en introduisant de nouvelles informations et de nouveaux processus dans la chaîne de synthèse de parole ;
- le traitement du langage naturel : le langage est présent à tous les niveaux en traitement de la parole de par les annotations nécessaires (texte, mots, phonèmes, style, prosodie, etc) ;
- la synthèse de la parole pour laquelle deux axes sont traités à travers les deux principaux types de systèmes (statistiques et par concaténation) ;
- l'évaluation des systèmes de synthèse de la parole.

Ces travaux se sont fortement élargis depuis ces dernières années avec l'aboutissement de la création d'un prototype de moteur de synthèse par sélection d'unités, la coordination du projet ANR Phorevox, de nouvelles collaborations et également l'obtention du financement d'un projet ANR JCJC, dont je suis le coordinateur.

Publications et production scientifique

Mon activité scientifique se traduit par 44 publications, dont 15 dans des conférences francophones à comité de lecture, 26 dans des conférences internationales à comité de lecture, 1 revue francophone à comité de lecture et 1 revue internationale à comité de lecture de rang A. Les deux conférences francophones dans lesquelles j'ai publié sont celles du domaine de la parole et du traitement du langage. Les autres publications ont principalement eu lieu dans des conférences internationales dont 18 dans des conférences de premier plan et 2 au workshop Blizzard associé au challenge international de synthèse de parole.

De plus, le développement de l'axe synthèse de parole dans mon équipe m'a amené à effectuer du développement qui a abouti au dépôt à l'APP de deux logiciels, dont un est distribué à la communauté.

Dépôt de logiciel auprès de l'APP :

- **Roots :** La librarie ROOTS (Rich Object-Oriented Transcription System) permet la représentation d'informations multi-niveaux organisées à travers des séquences dont les éléments peuvent être en relation. Elle a été développée dans le cadre de mes activités de recherche sur la thématique du traitement automatique de la parole, afin de garantir une représentation cohérente des annotations de corpus de parole pour diverses applications (notamment phonétisation, construction de voix, synthèse de la parole). ROOTS a été déposée en 2012 auprès de l'APP et ensuite rendue publique en 2014 [1]. La part de ma contribution à cette librairie est supérieure à 50% autant du point de vue des spécifications que du développement.
- **Moteur de synthèse par concaténation :** Le moteur de synthèse de l'équipe a été déposé auprès de l'APP en 2015. En raison de son caractère plus stratégique, il n'est pas diffusé auprès de la communauté pour le moment. Pour sa création, ma contribution s'élève à plus de 50% et concerne la spécification, la conception, le développement et les évolutions du système. Il est à noter que ce système nous a permis de participer au challenge international de synthèse de parole (*Blizzard Challenge*) en 2015 et 2016.

Publications majeures récentes

Mes publications se répartissent de manière équitable dans les thématiques précédemment citées. Parmi elles, certaines peuvent être retenues de manière particulière que ce soit pour leur importance vis-à-vis de futurs travaux en posant des briques indispensables, ou de l'amorçage de collaborations.

1. https://bitbucket.org/lolived/roots

[1] J. Chevelu, G. Lecorvé, and D. Lolive. ROOTS : a toolkit for easy, fast and consistent processing of large sequential annotated data collections, In Proceedings of the international conference on Language Resources and Evaluation (LREC), 2014.

[2] D. Guennec and D. Lolive. Unit selection cost function exploration using an A* based Text-to-Speech system, In Proceedings of the international conference on Text, Speech and Dialogue (TSD) 2014, 2014.

[3] M. Avanzi, G. Christodoulides, D. Lolive, E. Delais-Roussarie and N. Barbot. Towards the Adaptation of Prosodic Models for Expressive Text-To-Speech Synthesis, In Proceedings of the International Conference on Speech Communication and Technology (Interspeech), 2014.

[4] R. Qader, G. Lecorvé, D. Lolive and P. Sébillot. Probabilistic Speaker Pronunciation Adaptation for Spontaneous Speech Synthesis Using Linguistic Features, In Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP), 2015.

[5] G. Lecorvé and D. Lolive. Adaptative statistical utterance phonetization for french. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

[6] J. Chevelu, D. Lolive, S. Le Maguer and D. Guennec. How to Compare TTS Systems : A New Subjective Evaluation Methodology Focused on Differences, In Proceedings of the International Conference on Speech Communication and Technology (Interspeech), 2015.

[7] D. Guennec and D. Lolive. On the suitability of vocalic sandwiches in a corpus-based TTS engine, In Proceedings of the International Conference on Speech Communication and Technology (Interspeech), 2016.

[8] M. Tahon, R. Qader, G. Lecorvé and D. Lolive. Improving TTS with corpus-specific pronunciation adaptation, In Proceedings of the International Conference on Speech Communication and Technology (Interspeech), 2016.

[9] R. Qader, G. Lecorvé, D. Lolive and P. Sébillot. Ajout automatique de disfluences pour la synthèse de la parole spontanée : formalisation et preuve de concept, In Actes de la conférence TALN 2017, 2017. *(best paper award)*

[10] C. Fayet, A. Delhay, D. Lolive, P.-F. Marteau. Big Five vs. Prosodic Features as Cues to Detect Abnormality in SSPNET-Personality Corpus, In Proceedings of the International Conference on Speech Communication and Technology (Interspeech), 2017. (to appear)

Encadrements

Post-Doctorants et ingénieurs

- Jonathan Chevelu, IGR, 24 mois, 2012-2014, projet Phorevox, encadrement à 100%

- Sébastien Le Maguer, IGR, 3 mois, 2014, projet Phorevox, encadrement à 100%
- Marie Tahon, IGR, 21 mois, 2015-2017, projet SynPaFlex, encadrement à 50%
- Gaëlle Vidal, IGE, 12 mois, 2015-2016, projet SynPaFlex, encadrement à 100%

Doctorants

- David Guennec, 10/2012-09/2016, encadrement à 100%
- Raheel Qader, 01/2014-03/2017, co-encadrement à 30%
- Cédric Fayet, 11/2015-, co-encadrement à 37,5%
- Raphaël Pineau, 01/2016 - 05/2016, encadrement à 100%, démission du doctorant pour réaliser un projet de création d'entreprise
- Aghilas Sini, 12/2016-, co-encadrement à 50%
- Meysam Shamsi, 06/2017-, co-encadrement à 25%

Stages de master recherche

- David Guennec, 2012
- Cédric Fayet, 2015
- Sandy Aoun, 2016
- Antoine Perquin, 2017

Responsabilités scientifiques

Coordination de projets de recherche

Phorevox De juillet 2013 à fin novembre 2014, j'ai été coordinateur du projet Phorevox, en particulier en raison du départ du coordinateur précédent. Avant cela, j'avais participé à la construction du projet et à son suivi depuis son début en mai 2012. Ce projet financé par l'ANR vise l'application des technologies vocales pour l'apprentissage du français notamment en milieu scolaire et à destination d'un public de cycle 2. Il regroupe les partenaires suivants : l'Irisa/Université de Rennes 1, le Laboratoire de Linguistique Formelle (LLF), le Centre de Recherche sur l'Education, les Apprentissages et la Didactique (CREAD), la société Voxygen ainsi que la société Zeugmo. Il a permis de construire un prototype de plateforme en ligne reposant sur la technologie de Zeugmo concernant l'apprentissage du français, sur la brique de synthèse de parole de la société Voxygen, et des outils de création et d'annotation de corpus que nous possédons dans l'équipe. Des expérimentations en classe et auprès d'enseignants ont pu être conduites et ont permis de valider les différents éléments proposés dans le projet. Il a également été à l'origine de nouvelles collaborations qui se poursuivent à l'heure actuelle (notamment

avec E. Delais-Roussarie et M. Avanzi) et à l'élaboration un nouveau projet à l'ANR comme suite de ce dernier.

SynPaFlex En 2015, j'ai obtenu le financement du projet ANR JCJC SynPaFlex qui a débuté le 1er décembre 2015. Ce projet a pour objectif l'amélioration de l'expressivité des systèmes de synthèse de la parole en explorant deux axes de recherche complémentaires : la prise en compte des variantes de prononciation et la modélisation de la prosodie. Ce projet implique Elisabeth Delais-Roussarie, DR CNRS au Laboratoire de Linguistique Formelle, ainsi que Katarina Bartkova, MCF à l'université de Loraine et chercheuse dans le laboratoire ATILF. La durée de ce projet est de 42 mois avec un montant financé par l'ANR d'environ 250 k€.

Comités de sélection

- Comité de sélection, Université de Rennes 1, Enssat, 2012
- Comité de sélection, Université de Bretagne Sud, 2015
- Comité de sélection, Université de Rennes 1, Enssat, 2015

Jurys de thèse

- Thèse de Maël Pouget, Université de Grenoble-Alpes, soutenu le 23/06/17, examinateur.

Relecture et autres responsabilités

En terme de visibilité scientifique, je suis relecteur pour plusieurs conférences et journaux du domaine :

- Conférence francophone JEP - Journées d'Etude sur la Parole : conférence francophone de la communauté parole
- Conférence Internationale Interspeech : conférence internationale de référence pour le domaine
- Revue IEEE Transactions on Audio Speech and Language Processing
- Revue Traitement Automatique des Langues (TAL)
- Conférence Internationale ICASSP
- Conférence Internationale PAPE

J'assume également les responsabilités suivantes :

- Responsable de l'axe Parole de l'équipe Expression
- Membre élu au conseil scientifique de l'Enssat (depuis Octobre 2011)

- Membre élu du conseil d'administration de l'AFCP - Association Francophone de la Communication Parlée, depuis fin 2016

De plus, en 2015, j'ai effectué des expertises de projet de recherche pour l'ANR.

Diffusion des travaux (rayonnement et vulgarisation)

Lors de ma participation à la conférence JEP en 2014, nous avons participé à une session de démonstration à la fois tournée vers les chercheurs et le grand public. Le but était de présenter le projet Phorevox en conservant différents niveaux de compréhension, des problématiques scientifiques soulevées jusqu'aux enjeux applicatifs.

En terme de collaborations internationales, je suis en train de développer une collaboration avec Bernd Möbius et Ingmar Steiner (université de Sarrebruck, Allemagne) par le biais de Sébastien Le Maguer (ancien doctorant de l'équipe), autour des problématiques de synthèse de parole statistique et l'évaluation de la synthèse. Une autre collaboration est en cours avec Phil Garner (IDIAP, Suisse) en lien avec des travaux sur la prosodie. Ces deux collaborations se sont traduites par la préparation et la soumission d'articles ainsi qu'avec le séjour doctoral de David Guennec à l'IDIAP. Au niveau national, je collabore avec Elisabeth Delais-Roussarie (DR CNRS, LFF) depuis plusieurs années et désormais Katarina Bartkova (ATILF) avec le projet SynPaFlex. Une autre collaboration a démarré depuis novembre 2016, au niveau national, avec Anaïd Donabedian (Inalco) et Elisabeth Delais-Roussarie (LLF) pour la construction d'un moteur de synthèse de parole dans le cadre d'un projet de dictionnaire pour l'Arménien Occidental.

Enfin, en 2015 et en 2016, nous avons participé au challenge international de synthèse de parole *Blizzard*[2]. Chaque année, une dizaine d'équipes au monde participe à ce challenge. L'édition de 2015 concernait la synthèse de parole avec 6 langues indiennes (Bengali, Marathi, Hindi, Telugu, Malaylam et Tamil) tandis que l'édition 2016 était focalisée sur la synthèse de livres audio pour enfants en Anglais. En 2015, à l'issue du challenge, j'ai présenté le système de l'équipe lors d'un workshop à Berlin devant les spécialistes du domaine. De la même manière, j'ai présenté notre contribution en 2016 lors du workshop tenu à Cupertino, USA. Il s'agit des deux premières participations de l'IRISA à ce challenge.

Enseignement

J'effectue mes enseignements à l'Enssat, Université de Rennes 1, du niveau L3 au niveau M2, auprès d'élèves-ingénieurs en filière par apprentissage ou en formation initiale. J'ai en charge les enseignements suivants :

- Développement Orienté Objet, spécialité par apprentissage IMR, 1ère année (L3)

2. http://www.synsig.org/index.php/Blizzard_Challenge_2015

- Théorie des langages et Compilation, spécialité Informatique, 2ème année (M1)
- Traitement Automatique des Langues et de la Parole, spécialité Informatique et Master Recherche, 3ème année (M2)

Par ailleurs, j'ai eu en charge le suivi de 8 étudiants en formation par apprentissage tout au long de leur scolarité. Cette charge implique un suivi individuel des étudiants et des relations régulières avec les maîtres d'apprentissage.

Autres activités et responsabilités (pédagogiques, administratives)

De manière complémentaire, j'ai assumé une responsabilité d'année dès mon recrutement en tant que maître de conférences. En effet, une nouvelle formation par apprentissage est ouverte depuis septembre 2009 à l'Enssat. Cette formation s'oriente vers la spécialité Informatique, Multimédia et Réseaux. J'ai participé à la mise en place de cette formation et particulièrement celle de la deuxième année de formation (construction de la maquette d'enseignements, mise en place opérationnelle des enseignements). Par la suite, j'ai été **responsable de la deuxième année depuis la création de la filière et jusqu'en 2014**. Depuis septembre 2015, je suis **responsable de la 2e année de la formation d'ingénieurs en Informatique** de l'Enssat. Pour ses deux responsabilités, j'assume depuis mon recrutement une charge d'organisation des enseignements sur l'année, les relations avec les promotions d'étudiants concernées ainsi que les relations avec les maîtres d'apprentissage ou encore la gestion des stages obligatoires dans le cursus pour les élèves ingénieurs.

De plus, depuis octobre 2012, je suis également **membre élu au conseil d'école de l'Enssat**. Ce conseil est notamment en charge des grandes orientations prises par l'école.

J'ai été également **responsable par intérim du pôle Informatique** de septembre 2015 à janvier 2016. Cette dernière responsabilité, très consommatrice en temps et en énergie, m'a amené à gérer l'évolution des maquettes d'enseignement et leur mise en place ainsi que la participation à la direction de l'école. Ainsi, pendant cette période, j'ai participé à la réunion hebdomadaire de la cellule de direction de l'école ($\approx$ comité de pilotage) mais également à des réflexions sur les évolutions et la mise en place de la 3e année en Informatique pour la rentrée universitaire 2016.

Enfin, j'organise depuis 2014 le voyage pédagogique d'une dizaine de jours que doivent effectuer les élèves-ingénieurs par apprentissage de 2e année (environ 25 étudiants). Cela inclut notamment l'organisation d'un cours dans un établissement partenaire. Par exemple, en 2015 et 2016, j'ai organisé un cours délivré par John Kelleher, au *Dublin Institute of Technology*, en Irlande. Ces interactions ont également débouché sur la création d'un accord Erasmus entre nos deux établissements.

Annexe B

Projets de recherche

Cet annexe s'attache à presenter les travaux réalisés dans le cadre de projets de recherche collaborative. Notamment, les deux projets présentés sont le projet ANR Phorevox centré sur l'apprentissage de l'écrit dont j'ai été le coordinateur et le projet ANR JCJC SynPaFlex que je pilote actuellement.

B.1 Phorevox : apprentissage de l'écrit par l'usage de l'oral

Phorevox est un projet pluridisciplinaire dont l'objectif principal est de proposer un outil d'aide à l'apprentissage du français écrit par l'usage de technologies vocales. Les solutions technologiques d'aide à l'apprentissage des langues, et plus particulièrement du français, restent pour la plupart cantonnées à des exercices fondés majoritairement sur un usage de l'écrit (QCM, saisie de mots, pictogrammes). Nous pensons que l'interaction par l'oral peut conduire à une plus grande maîtrise de cet écrit. Pour rendre cela possible, le projet s'est orienté vers la création de contenus adaptés, allant de l'acquisition phonologique à des exercices plus traditionnels (dictée de mots, de phrases, production de phrases semi-libres), la création de voix de synthèse adaptées aux exercices réalisés, ainsi que la construction de profils d'apprenants pour une évaluation de leurs points forts et points faibles. Dans ce cadre, une plateforme en ligne orientée vers un public de cycle 2 (CP/CE1) a été créée avec un soin tout particulier à l'interface et aux retours effectués vers les élèves. Les résultats obtenus ont permis de lever la plupart des verrous énoncés dans le projet, à savoir la possible utilisation de la synthèse de parole dans un contexte éducatif, la restitution de styles prosodiques particuliers ainsi que le suivi des élèves avec la définition d'un profil rendant compte de leur niveau et de leur progrès. L'usage d'une telle plateforme permet de favoriser l'autonomie des élèves dans leur apprentissage en offrant un retour individuel et la gestion par les élèves de leur propre rythme de travail.

En particulier, une bonne part du travail réalisé dans le projet s'est orientée vers la construction automatique de dictées afin de les produire avec un découpage, un rythme

et une prosodie adéquats. Dans la suite, la problématique du projet est présentée de manière plus précise

B.1.1 Problématique et état de l'art

Le projet Phorevox, en proposant un système d'apprentissage du français écrit s'adaptant aux besoins des élèves, répond à deux demandes distinctes : l'une liée à la nécessité pour un grand nombre d'individus de mieux maîtriser le français écrit ; l'autre en relation avec la recherche de solutions pédagogiques innovantes permettant à chacun de travailler en autonomie. Dans la société actuelle, la maîtrise d'une langue, en particulier à l'écrit, est essentielle (échanges internationaux, mondialisation de l'économie, etc.). En France plus de 15% des élèves ne maîtrisent pas les compétences de base à la fin de la scolarité obligatoire, et 150000 jeunes sortent des écoles chaque année sans la moindre qualification. De fait, 20% des élèves sont en situation d'échec scolaire, chose très préoccupante. Une part importante de la population française reconnaît avoir des difficultés pour maîtriser le français écrit comme le montre un sondage réalisé par la société ZEUGMO auprès de 2500 personnes. Les applications dédiées à l'apprentissage du français pour un public FLM (français langue maternelle) ou FLE/FLS (français langue étrangère/ français langue seconde) sont nombreuses, mais parmi elles seul un petit nombre est spécialisé dans l'apprentissage de l'écrit (ex. Orthodidacte, Ortholud, Alphalire). À notre connaissance, seul Lectramini a recours à la génération de contenus (à partir de textes fournis par l'enseignant) et à la technologie vocale pour faire travailler le français écrit. Néanmoins, ce logiciel n'a pas les mêmes objectifs, se centrant davantage sur la lecture/compréhension.

Compte-tenu des relations entre représentations phonologiques et code écrit (Saussure de 1916 ; Catach, Gruaz et Duprez 1995), il est judicieux de toujours garder un contact avec l'oral lors de l'entrée dans l'écrit ou de tout passage à l'écrit. Aussi, la synthèse vocale peut être un outil didactique intéressant dans l'apprentissage de la langue (Carrier 2000 ; Marty 2005). Le potentiel de la synthèse de parole reste sous-exploité à l'heure actuelle (Gelan 2010). D'après (Maxine Eskenazi 2009), l'exigence d'une voix de synthèse de très haute qualité, indépendante du domaine et proche d'une voix naturelle pour servir d'exemple à l'apprenant a longtemps freiné le développement d'applications didactiques. D'après des travaux d'évaluation des produits commercialisés et utilisant la synthèse vocale (Zöe Handley 2009), il ressort qu'il n'existe pas de système de synthèse vocale universel pour l'apprentissage des langues. Dans ce contexte, une description précise des besoins devrait permettre la réalisation d'une synthèse de la parole (TTS) répondant aux critères d'intelligibilité, d'acceptabilité et d'appropriation à une application didactique dédiée à l'enseignement du FLM et du FLE/FLS. Cette étude confirme également l'attente des utilisateurs quant à une prosodie plus appropriée que ce soit pour des exercices d'apprentissage de la prononciation, de l'intonation ou l'utilisation de « systèmes de lecture » (dictionnaires sonores, dictées).

Parallèlement, la génération automatique de contenus et la définition d'un profil d'ap-

prenant doivent être traitées. Une difficulté est que ces deux problèmes sont très liés. Pour le premier, une approche possible est de fonder la génération de contenus sur l'utilisation d'un ensemble de textes. Ce dernier doit offrir une richesse linguistique suffisante. L'énoncé de l'exercice doit répondre à des contraintes précises, notamment en matière de longueur et d'évènements requis. Ce problème s'apparente au problème de couverture d'ensembles mis en œuvre dans la constitution de scripts de lecture (Chevelu, Barbot et al. 2008). Concernant la définition du profil d'apprenant, des informations dynamiques qui dépendent à la fois des compétences nécessaires à la réussite d'une tâche (segmentation, graphème/phonème, opposition phonologique), des difficultés ou troubles propres aux apprenants et aux résultats obtenus aux différents exercices, sont requises. Si les compétences attendues sont en général bien définies dans des référentiels, les erreurs que peuvent commettre les apprenants sont très variées et nécessitent l'expertise d'enseignants.

Deux principales évolutions sont à noter sur la durée du projet. La première concerne la restriction au public FLM en raison des différences importantes entre public FLE et élèves de cycle 2. La deuxième est la modification des priorités par rapport à la définition du profil d'apprenant et la génération d'exercices. Cela nous a conduit à travailler prioritairement sur la définition des caractéristiques du profil servant de base à la génération d'exercices.

B.1.2 Méthodologie

De manière générale, le projet se décompose en une succession de phases de spécification, de développement et d'évaluation. La spécification des exercices a donc été réalisée de manière prioritaire en collaboration avec des didacticiens (CREAD) et des enseignants intervenant en cycle 2. Cette première phase a permis de retenir trois types d'exercices à développer dans la suite du projet : les oppositions phonologiques, la segmentation en mots et la dictée. La construction d'une première version de la plateforme distribuée a pu être réalisée en parallèle. Celle-ci a nécessité un travail important de spécification des services rendus par chaque partenaire (plateforme Web par Zeugmo, service de synthèse vocale et de phonétisation par Voxygen, analyse des erreurs et de la progression de l'élève par l'IRISA).

De plus, la définition des formes des exercices (ergonomie, design) est un point essentiel en raison du public visé. Ce travail a été réalisé en collaboration étroite entre Zeugmo et le CREAD. Cela a permis l'implantation des différents exercices par Zeugmo, en prenant en compte leurs formes et en intervenant sur les bases de données nécessaires au stockage des informations liées à ces dernières.

Un point important mis en exergue par le contexte de l'apprentissage est le mécanisme d'autocorrection permettant à l'élève de progresser de manière autonome. Sa mise en place nécessite une analyse fine de la production de l'élève non seulement en utilisant des outils de correction orthographique mais également en comptant sur le fait que la

synthèse va pouvoir rendre les erreurs perceptibles. De plus, se cantonner à une validation unique de la réponse de l'élève n'apporte rien du point de vue pédagogique, ce qui nous a conduit à imaginer un mécanisme d'essai-erreur permettant à l'élève de se confronter à ses difficultés. De manière plus spécifique, le cas de la dictée a nécessité des études approfondies pour créer une voix adaptée, des modèles prosodiques reflétant la production d'un enseignant tout en se calant sur la vitesse de frappe de l'élève. Plusieurs problèmes se posent alors :

- Comment construire une voix « dictée » ? Quelles consignes appliquer lors de l'enregistrement ? Quel script d'enregistrement utiliser ?
- Comment reproduire la prosodie utilisée par un enseignant en classe pour une dictée ?
- Considérant la production de l'élève, comportant des erreurs souvent importantes, comment évaluer la position de l'élève dans la dictée par rapport à une référence ?

La résolution de ces problématiques a été réalisée de manière parallèle par les différents acteurs du projet et a conduit à un prototype évaluable par les élèves et les enseignants. Pour pallier le lien fort entre la génération automatique d'exercices et la définition d'un profil d'apprenant, nous avons fait le choix de mettre l'accent sur cette dernière tâche. Le profil est établi à partir des retours possibles du correcteur orthographique utilisé et de l'expertise des enseignants pour se ramener à une liste de critères acceptables pour des élèves de cycle 2. Pour la génération automatique d'exercices, un corpus de 1000 livres libres de droits a été construit et annoté automatiquement. Pour cela, une adaptation des structures de données existantes à l'IRISA a dû être effectuée. L'évaluation de la plateforme a été menée en plusieurs étapes au cours du projet. Ainsi trois phases d'évaluation se sont succédées :

1. Ergonomie/design de la plateforme, validation des contenus des exercices par les enseignants, première validation du principe de l'utilisation de la synthèse et retours sur le bénéfice de l'approche ;
2. Evaluation des corrections apportées (ergonomie et usage de la synthèse) ;
3. Evaluation de l'écriture guidée par une liste de mots auprès des élèves et du profil d'apprenant auprès d'enseignants.

De plus, des travaux préliminaires ont été menés sur les styles prosodiques et sur les voix avec emphase pour viser la lecture de poèmes et d'autres styles pouvant être utiles dans un contexte d'apprentissage des langues. Enfin, pour chacune des briques technologiques, des évaluations indépendantes puis combinées dans la plateforme ont été réalisées.

B.1.3 Résultats

À l'issue du projet, nous pouvons considérer que la finalité du projet est atteinte dans la mesure où nous avons construit une plateforme intégrant :

- La synthèse de la parole et des modèles de prosodie adaptés ;

FIGURE B.1 – *Interface de la plateforme dans le cas d'une dictée.*

- Des exercices conçus pour tirer partie de la synthèse vocale ;
- Un mécanisme d'autocorrection qui dépend de la situation de l'élève et de sa production tout en l'incitant à prendre du recul par rapport à cette dernière ;
- La collecte d'informations pour construire un profil avec un retour personnalisé effectué à l'enseignant.

Un exemple est donné sur la figure B.1 dans le cas de la dictée.

De plus, nous sommes en bonne voie pour permettre la génération automatique d'exercices. Les briques de base sont construites, il nous faudrait maintenant les assembler afin d'alimenter un parcours d'apprenant dépendant du profil.

Sur le plan de la réalisation des livrables, la grande majorité d'entre eux sont achevés tel que prévus à l'origine du projet. Seuls trois points ont dû être reconsidérés : l'évaluation initiale et le parcours, le public FLE et la génération d'exercices pour le parcours d'apprenant. Pour le premier point, l'évaluation initiale n'avait pas de sens avec des enfants (CP, CE1) car trop tôt dans l'apprentissage. Nous avons recentré la réflexion sur des tableaux de bord indiquant les types de faute réalisés par l'élève afin de construire un profil d'apprenant plus réaliste.

Nous pensons que le projet a permis des progrès significatifs de l'état de l'art que ce soit en TTS ou en didactique des langues. L'expérience acquise montre l'utilisabilité de la synthèse pour des applications nécessitant des adaptations prosodiques. De manière complémentaire, l'interaction écrit-oral pour l'apprentissage montre de sérieux avantages même si cela ne permet pas à l'élève de corriger toutes les erreurs (homophones par exemple).

Un des achèvements du projet par rapport à l'état de l'art de la TTS est l'applicabilité pour un style particulier. En effet, les progrès sur les modèles de prosodie et sur la technologie de synthèse permettent maintenant d'envisager d'autres applications (i.e., lecture de poèmes). Grâce à ce projet, de nouveaux enjeux apparaissent, notamment un élargissement du public permettrait une orientation vers l'aide à la rédaction pour un public adulte et/ou FLE.

B.1.4 Discussion

Les résultats obtenus rendent crédible l'utilisation de la synthèse vocale dans un contexte d'apprentissage bien ciblé comme celui adressé dans le projet. Cela constitue un progrès indéniable dans la mesure où cela répond à un verrou posé par le projet. De plus, la constitution de modèles prosodiques adaptés permettant d'agir sur la synthèse, et les études menées sur les styles prosodiques permettent d'envisager des résultats intéressants sur d'autres styles et dans un contexte applicatif plus large. Notamment, des travaux sur l'aide à la rédaction de textes pour un public adulte sont envisagés. La génération automatique d'exercices, bien que non implantée dans le système à l'heure actuelle, a permis de construire un corpus de taille importante qui pourra être versé à la communauté. Les outils de couverture de corpus ont été adaptés mais par manque de temps, leur utilisation dans le projet n'est pas complète. La définition d'un profil d'apprenant constituait le troisième verrou du projet. Différentes notions ont été abordées concernant le profil (informations sur l'utilisateur, sur les contenus, sur les réalisations de l'apprenant). Ces informations sont stockées pour construire des indicateurs performants pour l'enseignant via la description du profil d'apprentissage. Elles sont précieuses et constituent la base d'un parcours individualisé via un retour réalisé aux enseignants sous la forme d'un compte-rendu. Pour l'utilisation d'un tel profil dans une application grand public, il resterait encore à définir la notion de parcours afin de proposer des exercices automatiquement aux apprenants et ainsi établir le lien avec la génération d'exercices. Pour finir, les évaluations menées en classe montrent l'utilité d'une telle approche ainsi que le souhait de la part des enseignants de disposer d'outils adaptés.

B.2 SynPaFlex : flexibilité pour la synthèse de parole

De nos jours, la synthèse de la parole à partir du texte permet d'atteindre de très bons niveaux de qualité. L'usage de grands corpus de parole a pour une grande partie contribué à ce succès. Malgré tout, la parole synthétique générée manque encore d'émotion, d'intention et de style. À l'heure actuelle, nous ne sommes pas capables de synthétiser une voix comportant l'expressivité nécessaire pour de la lecture de livres audio sans enregistrer un locuteur afin de créer un corpus de grande taille possédant ce style.

Certains travaux dans le domaine s'intéressent à la prise en compte de phénomènes liés à l'expressivité et apportent des conclusions intéressantes permettant en partie de

caractériser le fonctionnement et la matérialisation de ces phénomènes. Nous envisageons ici de traiter de manière conjointe l'émotion, l'intention et le style d'élocution puisqu'en pratique ces notions sont très liées.

L'idée du projet SynPaFlex [1] est de s'intéresser aux différentes caractéristiques de ce qui fait l'expressivité d'une voix afin de constituer un modèle de prosodie et un modèle de modification de la chaîne phonémique, adaptés à un locuteur. Ensuite, l'utilisation de ces modèles sera explorée afin d'intégrer de l'expressivité dans les systèmes de synthèse de la parole par concaténation. Enfin, un complément de ce travail portera sur les post-traitements, nécessaires pour pallier les défauts des unités sélectionnées par le moteur de synthèse. Une approche par conversion de prosodie des unités sélectionnées sera alors envisagée comme post-traitement de la synthèse. L'ensemble de ces travaux portera sur l'étude du français et de l'anglais afin de conserver une certaine généricité.

Ces différentes étapes permettront d'apporter des connaissances sur la manière de modifier un système de synthèse, en terme de descripteurs des unités de parole, de fonction de coût de sélection des unités et de post-traitements.

Les enjeux majeurs du projet résident dans la faisabilité des applications de la synthèse de la parole expressive, applications qui pour l'instant restent peu répandues. Des débouchés sont notamment à attendre dans le domaine des jeux vidéo (diversification des voix de synthèse, création de voix expressives adaptées à la situation de jeu), de l'apprentissage des langues (dictée, style d'élocution) ou encore de l'assistance aux personnes.

B.2.1 Objectifs du projet

La vocation du projet est d'améliorer la flexibilité des systèmes de synthèse de la parole à partir du texte afin de générer une parole expressive de très haute qualité. Cet objectif principal passe naturellement par la réalisation d'objectifs secondaires :

1. la production de modèles du locuteur permettant de générer une prosodie particulière à un type de discours (haute qualité, adaptation à la tâche, adaptation au locuteur) ou à une émotion particulière (les deux aspects sont liés),
2. l'intégration de contraintes sur le processus de sélection des unités, en tenant notamment compte de critères liés à la réalisation de l'expressivité désirée,
3. et enfin la modification prosodique éventuelle des segments de parole choisis afin de satisfaire l'expressivité que l'on souhaite réaliser.

La réalisation de ces différents objectifs permettra de construire un système de bout en bout qui prend en compte de manière cohérente les différents types d'expressivité. Pour les atteindre, les principaux verrous que devra lever le projet sont :

- Comment décrire l'expressivité de manière symbolique afin de fournir des consignes en entrée d'un système TTS ? Ce premier point fait appel à deux aspects : la

1. http://synpaflex.irisa.fr

définition d'un ensemble de commandes d'entrée possibles afin d'avoir un contrôle externe sur le comportement de la synthèse mais également, et c'est là qu'est le verrou, l'extraction d'un contenu affectif, émotionnel à partir du texte. Dans ce domaine, les travaux sur l'analyse de sentiments ou d'opinions peuvent servir de base de travail. Un ensemble d'outils et de lexiques existent et peuvent être utilisés afin de générer des annotations pertinentes pour la synthèse. Également, une restriction à un ensemble de types d'expressivité semble nécessaire.

- Quels descripteurs pertinents permettent de caractériser l'expressivité de la voix ? On sait que plusieurs paramètres acoustiques sont modifiés en fonction du type d'expressivité dans le signal de parole. La fréquence fondamentale, l'intensité, les durées en sont quelques-uns. De nombreuses études existent dans la littérature pour différentes langues et avec des panels d'expressivité très différents. La problématique est ici de proposer des ensembles de descripteurs pertinents qui permettraient de différencier les différents types d'expressivité. De plus, le lien entre descripteurs symboliques et descripteurs acoustiques devra être établi afin d'autoriser leur prise en compte dans le processus de synthèse.

- Comment intégrer ces descripteurs dans la synthèse de parole par corpus ? Ce verrou concerne la flexibilité du système de synthèse en analysant et adaptant les différents étages du processus de synthèse. Il est notamment nécessaire d'adapter la chaîne phonétique, la structure accentuelle, les consignes prosodiques, la prédiction des phénomènes para-linguistiques telles les hésitations. De plus, la recherche même des unités pertinentes (sur le plan de l'expressivité) dans un corpus de parole doit être traitée ainsi que la modification des unités sélectionnées afin qu'elles reflètent l'expressivité voulue. Les consignes prosodiques générées permettront de guider le processus de sélection d'unités. L'enjeu réside ici dans la création d'un processus de guidage prenant en compte les consignes et le contenu du corpus utilisé pour produire la voix. Nous avons déjà engagé des travaux dans ce sens pour la gestion de la cohérence des durées phonémiques pendant la sélection d'unités.

- Comment modifier la voix d'un locuteur afin d'obtenir, à partir d'une voix neutre, une voix expressive ? D'une certaine manière ce dernier verrou rejoint le deuxième axe du verrou précédent. Il s'agit de proposer des modèles de transformation du signal de parole (ou des unités qui le composent) afin que le signal généré reflète l'expressivité voulue. Différents niveaux de difficulté peuvent être distingués ici. À une extrémité de l'échelle de difficulté, on peut définir un premier niveau pour lequel on dispose de toutes les informations dont aurait besoin un système de synthèse de la parole à partir du texte et à l'autre extrémité, ne considérer que le signal de parole observé sans information supplémentaire (à l'exception du type d'expressivité). On peut bien entendu définir différents niveaux de difficulté entre ces deux extrêmes en faisant intervenir des connaissances sur le contenu du message (segmentation en phonèmes, séquence des mots prononcés, étiquettes grammaticales, structure syntaxique de la phrase, structure prosodique, etc.).

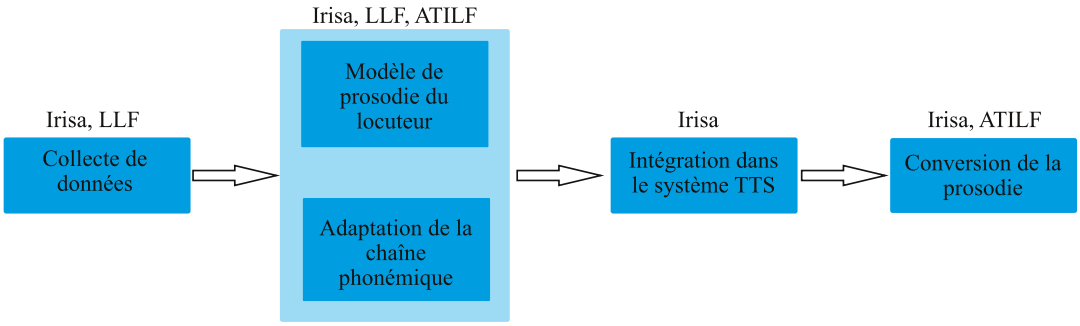

FIGURE B.2 – *Structure générale du projet SynPaFlex*

B.2.2 Organisation du projet

La structure générale du projet est représentée sur la figure B.2. De manière prioritaire, l'enregistrement et l'annotation de nouveaux corpus, avec les phonostyles et émotions traités, sera réalisé afin de compléter l'ensemble des données nécessaires à la réalisation du projet. Pour cela, nous nous plaçons dans le cadre de livres audio, permettant ainsi d'étudier les changements de styles (notamment passage au style direct) et nous nous focaliserons sur les émotions du *Big Six* (colère, dégoût, peur, joie, tristesse et surprise).

L'étude de l'expressivité (émotion, style, intention) sera la première tâche à effectuer afin d'extraire un ensemble de descripteurs pertinents en s'appuyant sur les nombreux travaux existant dans les domaines de l'analyse et de la reconnaissance de l'expressivité. La réalisation d'une bibliographie détaillée sur les travaux du domaine est nécessaire afin d'orienter les travaux ultérieurs. De cette étude, les éléments essentiels à la construction d'un modèle prosodique adapté aux phonostyles et émotions étudiés seront extraits. Leur pertinence sera étudiée sur les corpus dont nous disposons et ceux enregistrés dans le cadre du projet.

Les différentes briques utilisées dans les moteurs de synthèse doivent être adaptées : modification de la chaîne phonémique, prédiction de la prosodie, sélection d'unités, post-traitements après sélection des unités. Chaque brique donnera lieu à un ensemble de sous-tâches qui pourront être évaluées de manière indépendante les unes des autres.

Concernant la modification de la chaîne phonémique, il s'agit de prendre en compte explicitement les modifications (remplacements, élisions) de phonèmes lors de la réalisation de l'expressivité. En effet, lors d'une prise de parole, la chaîne phonémique réalisée est souvent différente de la séquence canonique que prédit un phonétiseur et porte une part de l'expressivité liée au contexte et au locuteur. Ces travaux sont complémentaires d'une thèse co-dirigée par Gwénolé Lecorvé et Damien Lolive sur la modélisation des variantes de prononciation en anglais, soutenue en mars 2017.

La construction du modèle prosodique est un enjeu majeur du projet puisqu'il doit permettre de produire, à partir d'informations linguistiques et de style, des consignes

prosodiques symboliques et numériques. Les consignes symboliques pourront être employées lors de la sélection d'unités tandis que les consignes numériques pourront être exploitées lors de l'étape de post-traitement après la sélection.

La sélection d'unités elle-même doit être modifiée de manière à prendre en compte des attributs pertinents pour l'expressivité tels que prédits par l'étape de génération de consignes prosodiques. Pour cette étape, plusieurs cas sont envisageables : des consignes symboliques (type d'expressivité, intensité, etc.) nécessitant un étiquetage préalable du corpus de synthèse (caractérisation des unités du corpus) ; des consignes numériques (contour de F0 de la syllabe en Hertz, durée des phonèmes en millisecondes, etc.) sont également utilisables afin de rechercher un profil particulier de segment acoustiques reflétant les propriétés voulues. Une difficulté est ici de savoir relâcher les contraintes de sélection de manière cohérente afin de préserver la qualité du signal en sortie.

Enfin, afin de capitaliser les connaissances acquises avec les premières tâches et explorer un nouveau domaine, l'étude de l'application de la conversion de voix comme un post-traitement à la synthèse sera réalisée. Cette étude permettra de s'intéresser à un problème moins difficile que celui de la conversion de voix inter-locuteur et pourra bénéficier de toutes les avancées réalisées précédemment.

Des évaluations seront menées au fil de l'eau de manière objective afin de mesurer l'avancement des différentes tâches et de communiquer des résultats intermédiaires à la communauté. De plus, dans chaque tâche sont identifiées des campagnes d'évaluations spécifiques qui permettront de prendre la mesure des résultats des travaux menés. Ces études incluront des évaluations perceptives qui seront menées sur la plateforme de tests subjectifs de l'équipe Expression.

B.2.3 Résultats

Le projet a débuté en Décembre 2015 et doit s'achever en fin Mai 2019. À l'heure actuelle, les tâches suivantes ont démarrées :

- Collecte et analyse des données : un corpus de livres audio d'environ 150h de parole a été collecté. Il est composé d'un sous-corpus de 80h enregistré par une locutrice et d'un sous-corpus de 70h enregistré par un locuteur. La totalité des données a été collectée à partir de Librivox, il s'agit donc de données libres qui permettront de partager le corpus analysé avec la communauté scientifique. L'analyse du corpus est effectuée en partie. Notamment, des annotation manuelles en terme d'expressivité selon des axes liés aux émotions ont été réalisées. Une analyse statistique du corpus est en cours. Cette tâche est réalisée par un ingénieur d'études et un doctorant recrutés pour le projet.
- Modification de la chaîne phonémique en fonction de l'expressivité : cette tâche est la plus avancée du projet en raison des travaux menés sur l'anglais pour la modélisation de la parole spontanée par Raheel Qader (thèse soutenue fin mars 2017). Une étude de l'adaptation de la prononciation en fonction du locuteur et de

la voix utilisée par le moteur de synthèse a été conduite. Les résultats montrent une amélioration à la fois de l'expressivité et de la qualité de la parole synthèse. Pour la réalisation de cette tâche, Marie Tahon a été recrutée en tant qu'ingénieur de recherche pour une durée de 21 mois.

- Construction d'un modèle de prosodie du locuteur adapté à l'expressivité : la construction de modèles prosodiques nécessite la finalisation du corpus de parole afin de disposer d'une quantité suffisante pour l'usage de techniques d'apprentissage automatique. Néanmoins, des travaux ont débuté sur l'analyse des phénomènes pouvant apparaître. Cette tâche doit être réalisée principalement par un doctorant recruté pour le projet et co-financé par le LABEX EFL.
- Intégration de l'expressivité dans les moteurs de synthèse de la parole : l'intégration de l'adaptation des modèles de prononciation à la chaîne de synthèse a été effectuée. La méthodologie mise en place permet un usage indépendant de la langue et repose sur des outils utilisés de manière standard dans l'équipe.

B.3 Conclusion

Dans cette partie, deux projets de recherche collaborative ont été présentés. Le premier, Phorevox, est un projet pluridisciplinaire dont l'objectif principal est de proposer un outil d'aide à l'apprentissage du français écrit par l'usage de technologies vocales. L'intégration des compétences de chacun des partenaires a permis de développer une plateforme en ligne proposant : des exercices adaptés à un public de cycle 2 ; un mécanisme d'autocorrection qui favorise l'autonomie ; une voix de synthèse adaptée aux exercices, notamment pour la dictée ; des modèles de prosodie permettant d'imiter le style d'un enseignant ; ainsi qu'un profil de compétences individualisé qui autorise un suivi de la part de l'enseignant. Les évaluations avec des classes volontaires permettent de conclure que l'approche est viable et apporte des bénéfices non négligeables à la fois pour les élèves et les enseignants. Également, de nouvelles pistes s'ouvrent désormais par rapport à l'aide à la rédaction pour les publics adulte et/ou FLE, ainsi que l'utilisation de cette technologie dans d'autres langues.

Le second est un projet qui porte sur l'amélioration du processus de synthèse de parole en vue d'améliorer sa flexibilité et le contrôle de l'expressivité. Deux axes principaux sont explorés par le projet : l'adaptation de la prononciation et la construction de modèles prosodiques, notamment pour la prise en compte des styles de parole. Les premiers résultats ont été obtenus et montrent que l'adaptation de la prononciation permet d'améliorer la qualité et l'expressivité de la parole synthétique. En termes de perspectives, ce projet permet d'envisager des suites sur la génération de styles de parole par transformation de voix (une tâche exploratoire est d'ailleurs présente dans le projet sur ce sujet), la prise en compte de modifications d'ordre lexical notamment par l'usage de techniques de type paraphrase.